Self-Knowledge and its Anomalies

Julie Germein

BA, MA (Hons)

This thesis is presented for the degree of
Doctor of Philosophy

Department of Philosophy
Macquarie University

May 2018

Acknowledgements

This work could not have been completed without the guidance and support of several people. Chief among them is my principal supervisor, Jeanette Kennett. Jeanette has been painstakingly patient and persevering in her help and support, both with the content of the thesis and with the structuring of my argumentation. At a late stage in my progress, she alone prevented me from falling apart. She also organised help from a visiting professor, Neil Levy, towards the end of my candidature. At that time I benefitted greatly from his assistance. Later, as I revised the thesis with Neil now as my official supervisor, he became the rock I could not have managed without. Thank you so much, Neil.

I benefitted earlier from the knowledge and experience of Catriona Mackenzie and Richard Menary, who both supervised the progress of my thesis at an earlier stage of its creation. Still earlier, I was helped and guided by my first supervisor, Albert Atkin, and his co-supervisor at that time, Cynthia Townley. Together they introduced me to the skills I would need to develop in order to write a PhD thesis.

Wendy Rogers supervised my first talk on self-deception in an early student seminar, and later chaired a discussion I had with Albert. I much appreciated her support at those times and later as well. Maryanne Hozijan also gave me much needed support at that time and this was also much appreciated.

Last but not least, I could not have survived the tumultuous and sometimes agonising experiences I have had in the writing of this thesis without the support of my children, Jem, Rosie and Ralph.

This thesis has been professionally edited by Dr Margaret Johnson of The Book Doctor, in accordance with the guidelines established by the Institute of Professional Editors and the Deans and Directors of Graduate Studies.

I certify that the work in this thesis entitled 'Self-knowledge and its Anomalies' has not previously been submitted for a degree; nor has it been submitted as part of requirements for a degree to any university or institution other than Macquarie University.

I also certify that the thesis is an original piece of research and it has been written by me. Any help and assistance that I have received in my research work and the preparation of the thesis itself have been appropriately acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Julie Germein

Student No. 69503818

April, 2017

Abstract

My thesis is that the practical irrationalities that we call anomalies are failures of the rational agent to know her own mind in situations where what she believes or intends or desires seems not to be rationally connected with how she acts. I take Richard Moran's 2001 account of self-knowledge as the basis for my argument that self-deception and akrasia, in particular, are better explained this way than as irrationalities unconnected with self-knowledge. Moran's account is written from the first-person perspective of the active agent, conceived as asymmetric with the third-person perspective. It sets out authority and estrangement as conceptual opposites. Estrangement, Moran argues, is usually caused by our 'burying' some thought that we find unacceptable: that is, by rendering it in some way unconscious. We can also be estranged from our attitude when we can think about it consciously, on Moran's account, by knowing it only theoretically, rather than immediately, for ourselves. Because she is estranged in some way from her real reason for her belief or intention, the subject suffers an unavoidable lack in her wholehearted knowledge of why she has the attitude, causing her to act with a degree of passivity concerning it. The passivity reduces her capacity actively to control her decision to believe that *p* or to do *a* on the basis of reasons. The consequence is that her self-ascription, although agential, is not fully self-determined. In both cases it has only third-person authority; she has failed to achieve normal, first-person authority over it.

Having defended Moran's account against deflationary objections, I then argue that estrangement renders all self-deceived self-ascriptions, all serious akratic self-ascriptions, and many everyday akratic self-ascriptions necessarily less than fully self-determined; that the relevant acts that follow are therefore necessarily to some degree psychologically unfree and thus that they have third-person rather than first-person authority.

Contents

Introduction

What could self-knowledge have to do with anything that we might want to call an anomaly?

An anomaly is something that puzzles, something that deviates from what is standard, normal or expected in some ordinary situation. We apply it to situations where what a subject believes or intends or desires seems not to be rationally connected with how she acts. Self-deception and akrasia, for example, are anomalous because they seem to violate a law-like generalisation that people's behaviour makes sense in the light of their beliefs. Where your self-ascription about your knowledge of your own mind (such as your belief or intention) is concerned, something anomalous about that self-ascription might be its inconsistency with how you act with regard to it. If you *sincerely* say that you intend to join the army but then immediately join your local choir instead and are unable to explain this satisfactorily, you seem to have deviated from an ordinary convention about the use of the word 'intention' that you would normally follow. If you announce that you are an atheist and then we find you attending church regularly, we might suppose that your sudden change of mind has been caused perhaps by excess emotion, such as guilt. Such anomalous situations are often attributed to character traits such as obstinacy, procrastination and imprudence, which also tend to produce irrational explanations of their actions by those who suffer from them. But most people, including philosophers, tend not to relate this irrationality to the subject's self-knowledge. They tend to discuss irrational self-ascription/action pairs from the point of view of practical irrationality.

My thesis takes a different view. Focusing on self-deception and akrasia as central examples of the phenomenon, it argues that these attitudes are

conceptually (rather than just causally) related to attitudinal[1] self-knowledge by being among its failures[2]: that is, insofar as we suffer from either of these anomalies we categorically lack self-knowledge to some degree because we suffer a relevant estranged attitude, i.e., an attitude that we *have* but are blocked off from being consciously aware of at that time (we cannot admit to ourselves, at that time, that we have this attitude).

The anomalies of self-deception and akrasia are, indeed, both practical and irrational, but to understand why they occur, I argue that we need to show how they are connected in this contradictory way with our knowledge of our own intentional attitudes. The anomalies confront us with the problem of unconscious thinking and feeling in many ordinary, daily situations, in ways that affect our self-knowledge and self-understanding as well as the way we behave. Self-deception, for example, may involve contradictions among a number of relevant attitudes, both conscious and unconscious. When the subject says 'I do not intend to go to work today', and the reason she gives to support this intention is that she is unwell, her real but unconscious reason may be that she is afraid of being confronted by her boss and having to explain why her last week's work is still unfinished. She cannot admit this to herself, however. The reason her last week's work is unfinished is her also estranged knowledge that last week she spent too much work time on the phone to her friends. She sincerely believes that she is unwell; indeed, she is not feeling too good because she also fears, consciously this time, that she might lose her job. But she is not sufficiently unwell to take sick leave. Being unwell is an excuse—

[1] This thesis discusses only the intentional attitudes, belief, desire, intention; and the affective attitudes, i.e., those that have an external object.

[2] The failure that estrangement brings is necessary, not contingent. Some might reasonably object to calling it a failure at all, since it necessarily prevents the subject from acting as she might if it were contingent.

a rationalisation. The scope of her anomalous self-deception covers all of these.

In akrasia, similar contradictions occur between conscious and unconscious reasons. You eat your apple pie and cream dessert having resolved not to. You believe you are weak-willed. But unconsciously, you eat because you feel depressed. (Also, you starved yourself all day in preparation for this dinner and therefore you now suffer low blood sugar.) Again, a whole range of feelings and thoughts, conscious and unconscious, as well as unknown physical events, are involved. You think that the explanation of your action is your weakness of will. But that's not the real explanation: rather, something else explains your apparent weakness.

My conviction that the anomalies are conceptually related to self-knowledge resonates with Richard Moran's *Authority and Estrangement: An Essay on Self-Knowledge,* published in 2001. For this reason, I use his account as the basis for this discussion, by setting it out, defending it and then referring to it from time to time to support my explanations of why I think self-deception and akrasia occur and what their conceptual links are with self-knowledge.

In his review of Moran's book, Sebastian Gardner (2004) describes what Moran is doing by saying:

> Richard Moran's endeavour … is to address the analytically circumscribed problem of the self in a way that undoes its isolation, restoring to its solution a connection both with the question of what we really and fundamentally are and with moral psychology, by reclaiming the thought that there is something 'distinctive or irreducible about the perspective of the first-person', that 'the person's own access and relation to his own mental life must be different in its possibilities and limitations from anyone else's.

Gardner describes Moran's epistemology as being substantive. By 'substantive', he means Moran's claim that attitudinal self-knowledge is

epistemic, representing a genuine cognitive achievement involving the awareness of some independently obtaining [mental] state of affairs and as having truth-conditions that are independent of the subject's making of the judgment.[3] This contrasts with what Moran calls the deflationist view that first-person self-ascriptions [4] are not expressive of substantive, first-person judgments.

The core of Moran's account is the claim that the first-person position is practical and deliberative (Gardner, p. 251) rather than theoretical, as it is in the third-person perspective. Moran's primary commitment is to discussing the differences between the first-person and the third-person perspectives, where the first, he claims, is irreducible to the third with respect to the intentional attitudes. This alone, if his account is viable, enables him to claim that the self is philosophically connected i) with the question of what we really and fundamentally are, and ii) with moral psychology. In this way, Moran constructs a philosophy of self-knowledge that connects it with some of its deficiencies is important. These deficiencies, Moran argues, include the fact that we can 'bury' certain things about our own mental lives that we do not wish to know consciously, thus producing what I am calling the anomalies.

With Moran, I believe that estranged attitudes are commonplace in our lives, that they do produce the anomalies, and that the misunderstandings these anomalies cause wreak havoc in many ways. We have become very good at developing new technologies to deal with our external needs. We all hope that this will continue. But sooner or later, I believe we also need to look within ourselves at what really drives us, and accept it and learn to live with it; our

[3] Gardner p. 250, referring to Moran 2001, p. 13. This is how I also use the word 'substantive' in this thesis.

[4] Self-ascriptions that have first-person authority are referred to from now on as avowals.

methods of killing each other are becoming too lethal. My reason for adopting Moran's account is not that it is anti-deflationary but rather that it focuses on the unconscious forces that produce estrangement.

The concepts of authority and estrangement as Moran explicates them in his book are basic to my thesis. The authority we give to our claims about our intentional mental states has long been recognised as special because they are knowable immediately. When we declare our belief that it is raining or that we feel happy or that Jones is still on holiday, we are not expected to provide any evidence in support of our claim to this belief. A belief known immediately might be false ('I believe today is Tuesday' [said on Monday]) or irrational ('I believe it will rain tomorrow because otherwise my plants will die.'). However, we can know immediately (groundlessly, without evidence or inference) that we *do have* the belief that today is Tuesday or that it will rain tomorrow. This immediacy makes our knowledge of our own intentional attitudes different from our knowledge of almost anything else. We can usually know such things as where our foot is without evidence, but not much more. However, Moran claims more than just immediacy for such 'I' statements. He argues that the first-person perspective on self-knowledge, the stance from which the subject says 'I believe that *p*', is different in kind from and asymmetric with the third-person perspective from which the subject says, 'Jones believes that *p*', because from the first-person perspective the subject can avow his belief while from the third-person perspective he can only attribute it to himself. The differences between these will be spelled out shortly.

But there is a problem both with immediacy itself and with its presumed special authority. Let us consider immediacy first. Since we are all alike in being human beings, how could there be any such irreducible difference between our knowledge of our own attitudes and our knowledge of the attitudes of others? How could there be any kind of self-knowledge, in fact, that does not involve

access to evidence that is available, at least in principle, to others as well? 'Claims not based on evidence do not in general carry more authority than claims that are based on evidence, nor are they more apt to be correct' (Davidson 1984, p. 103). What can the 'immediacy' of our knowledge of our own attitudes mean, unless it means something that must be explicable in some way that is objectively consistent with other kinds of knowledge? And since evidence and inference are necessary elsewhere, how could they not be necessary here? Surely, observation of evidence—perception of it, at the very least—must be 'in the mix' somehow. Although Moran dampens down the infallibility of the historical Cartesian observational picture by admitting to fallibility and error *in* immediacy, this seems to leave the problem of immediacy itself unresolved. If we can 'read' our own minds, for example (whatever this might mean), could not someone else, given the right equipment, be able to 'read' them as well?[5]

So immediacy is a puzzling notion to many philosophers, who may describe it as 'mysterious' (Fernandez, 2003, p. 352). It is not surprising that some accounts of self-knowledge insist that self-knowledge from the first-person perspective, although we seem to have privileged access to it, 'requires for its justification only the same conceptual elements already needed to explain perceptual knowledge' (p. 352). This is a strong claim, and requires an explanation of why I adopt Moran's account of self-knowledge to explain the anomalies as I do.

Moran does not consider the person neuroscientifically but rather in a commonsense, practical, folk-psychology way, as a rational agent who can think and act for herself. We do tend to say that when something causes

[5] Moran (2001, p. 91) suggests that even if we could 'read our own minds', we would still have to decide whether we agreed (or ever did agree) with what we 'found' there, since in mind-reading we would be considering ourselves evidentially, in a third-person way.

someone to become forgetful or confused, it interferes with her agency. We do not usually give a cognitive scientific explanation of this. We say that it is because the person, as agent, wants to switch on the TV that she gets up, finds her remote and switches on the TV. We think that she, as agent, is normally in charge of what she does.

Moran is not the only philosopher who believes that the folk-psychological approach is best to take on first-person authority. Davidson, for example, says of immediacy:

> The speaker can be wrong about what his own words mean. This is one of the reasons first-person authority is not completely authoritarian. But the possibility of error does not eliminate the asymmetry. The asymmetry rests on the fact that the interpreter [listener] must, while the speaker doesn't, rely on what, if it were made explicit, would be a difficult inference in interpreting the speaker. (1984, p. 110)

Here, Davidson is pointing out that the listener has no way of using evidence to assess whether the speaker is right or wrong. Even if *what* the speaker believes is false, and the listener knows this, *that* he believes it must normally be assumed. Even if the listener is aware that on other occasions the speaker has denied what she is now claiming, he cannot be sure that what he is hearing now is not the case at this point. When I object to the deflationary claim that 'immediacy' is really inferential, I do not mean to downplay deflationism as it is importantly used in the service of neuroscience; rather, I am saying that first-person self-knowledge has properties that third-person knowledge lacks. I defend Moran's account against deflationary accounts in order to use it to explain the anomalies.

When immediacy in self-knowledge is accepted, differences between the first-person and the third-person perspectives can emerge. In the first-person, avowable perspective, which Moran argues is the more fundamental of the

two, the subject reflects on whether *p* not only as an agent who might be just reporting her belief but also in a practical way that involves her own commitment to the *truth* of what she is avowing. This truth is about whether *the world* 'agrees' with her avowal. This does not mean that she has a special authority *over* the truth of what she believes. *What* she avows she believes may be false; the world may not 'agree' with it. She has made up her own mind to believe it nonetheless. The special authority of its epistemic privilege is built on this fact.

The other kind of 'I' statement, a kind which Moran claims is asymmetric with an avowal, is an attribution. Although both kinds of self-ascription have the authority of immediacy, an attribution does not also have the special authority of an avowal. In an attribution, the subject does not speak for her belief. She just says that she has it. (Moran, 2001, p. 93) A subject might say 'I believe that *p*' simply because she has been asked whether she does. Her intention might be to self-attribute the belief that *p* just to answer the question.

However, in such cases she may or may not be able to justify her self-ascription if asked. Moran argues that for a self-ascription to have the special authority of an avowal the agent must be able to justify it if asked, or, at least that it must be answerable to justification. Although both attributions and avowals are epistemic, only avowal is also practical. The avower speaks for her belief, stands up for it. Even if others think her belief is false, she believes it is true for a reason to which she is answerable.

This means that the kind of authority the subject has over her knowledge of her belief is linked with whether she can make a good fist at justifying her belief, or whether she can only self-ascribe it in a third-person, empirical way. Moran flags the difference by arguing that to be justifiable, an avowal must conform to the transparency condition: that the subject can explain what it is about the *object* of her avowal that she thinks makes it true. Her explanation

cannot be about herself as avower, such as where she learned it. Transparency is discussed in Chapter 1 and defended against objections to it in Chapter 3. First-person authority assumes that the subject's self-ascription is based on reasons that are minimally rational and that she can use, if asked, to *justify* her self-ascriptions because she has formed them *for herself*: it respects the reasoning power of the subject *qua* agent. In millions of daily cases, this respect is appropriate. But not all attempts at avowal can be justified. The presumption that an avowal is special because it is knowable immediately is just that—a presumption—and its usefulness can be and sometimes should be discounted in practice. If the subject cannot justify it 'by reflection on its subject matter alone and not by consideration of the psychological evidence for a particular belief attribution' (Moran 2001, p. 84), we must say that it is not an avowal at all but rather an attribution.

First-person authority is also linked with immediacy on Moran's account in his claim that when the subject believes that *p* she can *know* immediately that she believes that *p*. This has also attracted objections: how can the step from believing that *p* to knowing that you believe that *p* be immediate? The common sense, folk psychological position on this matter is that the knowing begins at the very beginning. When you are asked or ask yourself whether you believe that *p*, you may investigate this possibility intentionally and therefore knowingly. When you decide that you judge that *p*, you may therefore make *that* decision knowingly, too. By then, therefore, you already know that you judge that *p*. No further step, no inference, is required to take you either from judging to believing or from believing to knowing that you believe that *p*. As long as you are certain that you do judge that *p*, your own conviction now is that you believe that *p*. So now you know that you believe that *p*. I argue in Chapter 4, via Boyle's reflectivism, that your knowing that you believe that *p* is immediate. Boyle gives us philosophical rather than only common sense reasons for this claim.

But even if first-person authority is both immediate *and* asymmetrical with attribution, how can its asymmetry be even relevant to the anomalies of self-knowledge, let alone central to them, as this thesis argues?

My answer lies in the phenomenon of estrangement. I argue that all cases of self-deception and serious cases of akrasia necessarily involve the subject's estrangement (alienation) from her belief or action because at that time she necessarily cannot know, consciously, immediately, wholeheartedly and in a settled way, *why* she believes that *p* or intends to *a*. She does know, of course, *that* she believes that *p* (the serial gambler may really believe he will not gamble that night) but believing only *that p* is insufficient for having first-person authority over it if you cannot say, in a justificatory way, *why* you believe it.

In serious cases, estrangement is more likely to be due to 'Freudian repression', and in everyday cases it is often due to alienation from both alternatives owing to unresolved conflict concerning them. In Chapter 1, I argue that in the structural relation between authority and estrangement that Moran's account gives us, estrangement necessarily prevents an avowal from succeeding by causing a degree of psychological lack of freedom in the subject. The conceptual relation between the anomalies and self-knowledge gives us the key to explaining the anomalies of self-deception and akrasia via their connection with self-knowledge.

I use the phrase 'Freudian repression' as an example of a kind of estrangement only to reflect the common sense notion at large in society today that Freud talked about unconscious ideas. I do not wish to imply anything about Freud's theory of the unconscious mind, in particular his claim that the mind is divided into conscious and unconscious 'systems'. My own view is that much of what he said has long been superseded, although I think that the notion of unconscious mental states, along with some of his other claims, still remains

and is important. Nor do I wish to consider discussions of Freudian theory[6] that are in harmony with Freud's claims about topographic or qualitative 'regions' of the mind. Agnes Petocz (1999), for example, has disabused us of the idea that unconscious mental states and conscious mental states belong in different 'regions' of the mind. Where repression is concerned, for example, she points out that on Freud's 'structuralist' account of the mind, 'the repressed is repressed *because* it is unconscious' (p. 153, author's italics). I agree with her that that idea gets things back to front. In my view, unconscious mental states may be unconscious because they have been repressed.

Basing my account of the anomalies on Moran's account of self-knowledge reduces my thesis to manageable proportions. In general, I argue that estrangement, as just defined, gives us a more satisfactory way of explaining the anomalies than contemporary orthodox accounts do, by demonstrating that the weakness that anomalous acts are said to involve may be better understood as a lack of fully active agency than as a weakness of willpower. Estrangement, Moran argues, is caused by our 'burying' some thought that we find unacceptable by rendering it in some way unconscious. We can repress it, to use Freud's term, or avoid it less severely by temporarily 'blocking off' from thinking about it or by simply not being able or not knowing how to think about it at that time. We can also be estranged from our intentional attitude even when we can think about it consciously, on Moran's account, by feeling conflicted rather than wholehearted about it. I mostly use 'estranged' to refer to such attitudes when they are unconscious because of having been repressed. If it is explicitly conscious, then either the reason the subject has it

[6] Davidson's 1982 theory of the divided or partitioned mind, for example, seems to me to be unhelpful.

is inaccessible to her conscious awareness or she knows this reason consciously but only theoretically—its phenomenological immediacy is inaccessible to her (she is ignorant of how this immediacy would feel).[7] I mostly use 'alienated' to refer to attitudes estranged in this conscious way. In all of these situations, the subject's action is thus never fully *hers*; she cannot fully and sincerely endorse it. For this reason, her self-ascription, although *agential*, is to some degree psychologically unfree.[8]

The literature acknowledges that self-deception has its home in folk psychology, and thus may not be amenable to strict definition (Bayne & Fernandez 2009, p 1). The apparent intentionality of self-deception is also hard to explain as practical irrationality, since if you intentionally deceive yourself you know that you are doing so, and so it would seem you are not deceived. I argue that standard accounts of self-deception such as Mele's (2009), that claim that self-deception is based on biasing motives, cannot satisfactorily explain how such bias causes self-deception unless the subject is estranged from her real reason for her biased belief: otherwise, she is not self-deceived. One problem with akrasia (often defined as weakness of will) is that it is hard to differentiate it from compulsion. The following is an attempt to differentiate it:

> In cases of compulsion … the compelled agent lacks the rational capacity to bring her desires into line with her belief about what she ought to do, whereas in cases of weakness of will she has that capacity, although she does not exercise it (Stroud & Tappolet 2007, 10).

---

[7] This is a bit like Socrates' and Aristotle's claims about akrasia that sensory experience can convince us to *a* more thoroughly than a theoretical reason can convince us to ~*a.* This gives us another reason for preferring an account based on rationality to one based on cognitive science.

[8] The phrase 'psychologically unfree' will be elaborated by examples throughout.

But why does the akratic subject not exercise this capacity in cases where she is *not* compelled? I treat serious cases of akrasia differently from everyday cases. I argue that in serious cases, it is impossible to act knowingly and completely freely against your better judgment all things considered. If you do so act, therefore, you do so in a way that is to some degree psychologically unfree. In everyday akrasia, subjects are also psychologically unfree to some degree. These subjects are more likely to exhibit reckless or thoughtless behaviour, and in some cases this would involve a degree of alienation (estrangement). In such cases, their knowledge is psychologically unfree to some degree simply because they suffer conflict. You might be somewhat alienated from your intention both to *a* and to *~a* because you are conflicted about which of them to do. You cannot do either of them wholeheartedly. Your act is akratic because it is not completely self-determined.

Moran's account of self-knowledge has been widely criticised. In order to use it as I do, I must therefore first set out his account and defend it before applying it to self-deception and akrasia. It seems to me that to explain the anomalies, we need to use what I take to be a fact: the first-person perspective on our knowledge of our own intentional attitudes is asymmetric with the third. The very fact that we take self-deception and akrasia to be anomalies at all is based on our finding them too irrational for us to be able to explain them satisfactorily.[9]

The concepts of first-person authority, rationality, substantivity, transparency, epistemacy[10] and estrangement are basic to Moran's account. He argues that

---

[9] This assumes that current accounts of these anomalies are not completely satisfactory. I argue this way in Chapters 5 and 6.

[10] Epistemacy is about knowledge. For example, we can see objects without knowing what they are, but we cannot see them epistemically without knowing *that* what we see is (say) a cat. Epistemacy is closely associated with cognition, which is about thinking, as in Descartes' *Cogito*.

these fit together in ways that in fact *imply* irreducibility. I have already introduced the first of these, first-person authority, and the last, estrangement. First-person authority depends on the second, rationality, because it is via reasoning that the subject makes up her own mind whether she believes that *p* or desires or intends to *a*. On Moran's account, where *p* is a proposition about the external world, the first-person authority of our belief that *p* comes from the fact that as active, rational agents, we use our reasoning capacity as it is at the time to decide whether we think *p* is true. In deciding that it is true, we are deciding that we believe that *p*. This (usually but not always reflective[11]) process depends, in turn, on the third major concept in Moran's structure, transparency. Moran uses the phenomenon of transparency to explain how rational reflection can determine not only whether the subject believes that *p* but also whether *p* is true. Discussions of transparency begin with a suggestion by Evans, among others, that when we are considering whether to believe that *p* we focus on the external object of *p* rather than on our own minds. By considering the evidence we have, we make (form) our belief for ourselves by making up our own minds whether *p* is true; in deciding that it is true, we *are* deciding to believe that *p*, unless our decision is uncertain or unsettled. When it is certain and settled, it is avowable as long as it is minimally rational and conforms to transparency. The authority of our intentional attitudes, I will argue, is thus *first-personal* and both subjective and objective. Because we can make (create) our belief for ourselves, as active agents[12], knowing what we are doing as we do so, we can know what we believe.[13] So we can know our own minds about whether *p*. But we cannot form just any belief that *p* because our

[11] Not all acts of making up one's mind involve reflection. 'It's raining outside' usually needs no further comment to be acceptable.

[12] This is contentious, as there is a large literature arguing the opposite: that belief formation is not under our agential control. I argue in Chapter 3 that this other approach fails.

[13] In Chapter 4, I present Boyle's supporting position on *how* we can know what we believe.

rational agency is constrained by the external facts we are considering under the transparency condition; we cannot decide, at whim, that *p*—the world must seem to us to 'agree' with the conclusion of our reasoning. Our reflection can amount to our *knowledge* of the belief we form about its object, rather than being just some guess about this, only because our forming this belief *is* so constrained by external facts; the beliefs we form are assumed to be determined by our consideration of these facts. We can know that we believe that *p* or intend to *a* because we can know that *p*, as we see it, is true or that what we hope to achieve via our desire or intention to *a*, as we see it, is objectively choice-worthy. We can know *this* because we have formed our belief that *p* or our intention to *a*, sometimes via much reflection, on that very basis. We can normally express this knowledge, therefore, with first-person authority whenever our beliefs and intentions, as we see them, are objectively viable also in the judgment of relevant others. (This is to assume that such mental states are substantive and epistemic.) When others do not agree that our belief is objectively true, our authority has been damaged, either empirically, when our thinking is caused by (say) fatigue, or necessarily, by an estrangement over which, at that time, we have no control. In the former case, the belief is probably not settled or can be corrected later. In the latter case, the estrangement necessarily creates a degree of psychological unfreedom in the avower.

This brings us to epistemacy and substantivity. When a subject realises that she believes that *p*, she may either attribute it to herself or avow it. If she avows it, the first-person authority of her self-knowledge is therefore normally both epistemic and epistemic-plus. The knowledge is epistemic because it is a

substantive[14] cognitive achievement that can be true or false, both when she can only attribute it to herself and when she can avow it. The 'plus' in the stance of avowal indicates the avower's practical endorsement of and commitment to its truth or choice-worthiness. Both are required for the first-person authority of immediate self-knowledge. The authority of intentional attitudes self-ascribed from the first-person stance is special because it is immediate, because the subject has made up her own mind about it, because it is normally justifiable and because it is minimally rational. Moreover, its rationality is based on facts about the external world that the subject is using, implying, therefore, that she takes these facts to be true or false, depending on whether and how the independently existing world co-operates with her intention. Thus, she and her listeners treat her belief about her intention as being also *substantively* true.

When, in deflationary accounts, the difference between attribution and avowal disappears, so do practical endorsement and commitment.[15] Rational agency is normative because using reasons implies working out, on the basis of evidence, what one ought to believe, desire, feel about something or intend to do[16] as a preliminary to deciding what *to* do. (It leaves open the possibility that you might still decide that although you ought to mow the lawn, you will go to a movie instead.) Thus, Moran argues that the first-person perspective is asymmetric with the third-person perspective.

Moran's final major concept is estrangement. On many occasions we cannot

---

[14] 'Substantive' means 'having a separate and independent existence, not merely inferential or implicit', or 'the genuine "detection" of some independent psychological fact' (Moran 2001, p. 13). It is discussed in Chapters 1 and 4.

[15] This removes the normative dimension of any rational agency claim, reducing first-person justification to third-person explanation.

[16] Moran's account of Kantian rational agency is set out in Chapter 1 and defended in Chapter 2.

achieve a satisfactory degree of rationality. When the subject is estranged in some way from her real reason for her belief or intention, she suffers an unavoidable lack in her wholehearted knowledge of why she has the attitude, causing her to act with a degree of passivity concerning it. The passivity reduces her capacity actively to decide to believe that *p* or to do *a* on the basis of reasons. The consequence is that her self-ascription, although still agential, is not fully self-determined. In making estrangement central to the explanation of those cases where self-knowledge does not have ordinary, practical, fully active, first-person authority, the account thus leads us precisely to the theoretical space where the anomalies, such as self-deception and akrasia, can be found. An anomalous self-ascription is agential, but this agent's attempt at expressing her fully first-person authority over it has failed on this occasion.

The irreducibility thesis enables us to distinguish between self-knowledge that is normative (from the first-person perspective) because the subject can successfully endorse it as true (for beliefs) or objectively choice-worthy (for the other intentional attitudes—intentions, desires and the affective attitudes) and self-knowledge that is *only* epistemic, being no more than a third-person statement of psychological fact about what intentional attitude the subject *has*, where its truth or choice-worthiness is irrelevant. His account thus introduces us to the fact that third-person, deflationary accounts of self-knowledge cannot distinguish between the first-person, avowable perspective and the third-person, empirical, attributive perspective, because they cannot allow that there is an irreducible difference between the two. But although both perspectives are agential, both involve the subject's knowledge that she believes that *p* or intends to *a* and both are epistemic, only the first-person, avowable, practical perspective is normative as well as and includes the subject's endorsement and support of the truth of what she has avowed.

So I approach my topic from the point of view that practical, first-person self-

knowledge of our intentional attitudes is asymmetric with third-person self-knowledge. Subjects who successfully avow some belief, desire, intention or affective attitude are functioning as fully active agents at that time. To support my position, I set out Moran's claim that the agential basis of first-person, avowable self-knowledge lies in Kantian active, rational agency. His account draws on Kant's distinction between pure apperception, signified by the 'I think', and empirical apperception or inner sense, where the agency involved is more passive, as in self-ascribing a sensation. For Moran, as for Kant, pure apperception is the more fundamental. A subject's deliberation whether *p* involves active agency because she forms her belief on the basis of her own reasons, for better or for worse, and so the belief she forms is as rational as her belief-forming process allows on that occasion.

Having explored the major concepts in Moran's account in Chapter 1, I defend it in the next two chapters against objections to two of its major ideas: i) its basis in Kantian rational agency and ii) its explanatory focus on transparency. The objections I defend against are mainly attempts to reduce first-person authority to third-person authority.[17] Thus in defending against deflationist objections to these two ideas, I also defend the account's basic claim to

[17] A note on terminology: in using 'authority' in this way, I depart somewhat from Moran, who uses the term to indicate an active role of the person in the production of the belief. While I agree with Moran that first-person authority entails such activity, third-person authority consists only in coming to know a belief in a way that may set the stage for taking responsibility for it. Coming to know our estranged attitudes may require third-personal stance, but this is an essential stage on the way to making them our own.

asymmetry.

In Chapter 2 I answer objections to Moran's account of rational agency. Two standard criticisms of this are that it is too rationalistic and that Moran's concept of first-person avowal is reducible to that of third-person attribution. Owens (2003) and Carman (2003), for example, have both criticised it for being too rationalist, and Owens has also attempted to reduce first-person avowal to third-person attribution in the case of intentions.

I argue that Owens fails to make either case. His basic mistake is to suppose, as does the traditional, deflationary view of akrasia, that deliberately irrational akratic acts can be acts of free agency *par excellence*. I argue that in such cases the subject is conflicted between what she knows is against her better judgment all things considered and what she does. Conflicted thinking cannot be fully actively self-determined—the conflict produces a degree of passivity in such thinking. This does not necessarily imply that the subject suffers an estranged (repressed) attitude unless the case is serious. But it does imply a degree of conscious or unconscious alienation in the subject. Following this, I argue that Carman's objection that Moran's account is too rationalistic also fails. Carman claims that to have first-person authority on Moran's account, intentional attitudes must be justifiable by 'cognitively articulated' reflection or deliberation. I argue that answerability to reasons as Moran explicates it requires current cognitive deliberation only in some cases.

In Chapter 3 I defend Moran against objections to his transparency account. I consider two major objections: i) that transparent self-knowledge does not apply to pre-existing beliefs and ii) that transparency cannot explain immediacy. Both objections aim to reduce active agency to passive agency and first-person, reasons-based authority to third-person, attributive authority.

In Chapter 4 I show how the conscious belief that *p* can co-exist in the subject's

mind with the unconscious belief that not ~*p* as long as one of these beliefs remains unconscious at those times when a self-ascription of the other is conscious. I set out Boyle's (2011) concept of reflectivism and then discuss questions that arise from it. I argue that estranged attitudes are conceptually linked with rational agency because, unconsciously, they 'tell' the subject what she ought to believe or do. Does a certain mother's daughter suffer from learning problems? The mother cannot accept this possibility; it is too shameful. But unconsciously she knows it is true. At the time, she can know this only tacitly while the knowledge remains estranged. Once explicitly recognised, however, such attitudes must take their place as the real reason for the subject's belief, desire, affective attitude or intention. When this occurs, the subject now has a wider spread of immediately and rationally known attitudes; thus her agency is enhanced. She can now act with more confidence concerning her new self-knowledge, and because she no longer has to use up energy in continuing to repress what she has just brought to conscious awareness, she may feel more alive and have more energy than before. Moran claims that his account reinstates the connection between i) the irreducibility of the first-person perspective on self-knowledge to the third, ii) the moral dimension of the first-person perspective and iii) the connection between ii) and the subject's ongoing psychological well-being.

In Chapter 5 I discuss self-deception. I place Moran's account of self-knowledge in the self-deception literature, arguing that it can give us a new, useful perspective on self-deception: this is that self-deception is always caused by an estranged or alienated attitude. I support this position with examples.

In Chapter 6 I discuss akrasia. I argue that the seriously akratic subject suffers an estranged attitude, and to support my position on the relation between self-knowledge and the anomalies more generally, I discuss an example of akrasia

used by Christine Tappolet (2007), arguing that because hers is a third-person account it cannot differentiate between a subject's first-person and third-person attitudinal stances.

By the end of Chapter 6 I have set out Moran's account of self-knowledge, defended it against deflationist objections, and used it to show that self-deception and akrasia are better explained by an account of self-knowledge that does not reduce the first-person perspective to the third. My account makes the anomalies of self-knowledge no longer anomalous; they have been integrated into an account of self-knowledge by being placed among the contradictions and conflicts that a subject suffers when estrangement (alienation) is present. The explanation I give is congruent with our supposing that our mental processes are not just at the whim of an inexplicable disposition to act against our own explicitly known better judgment, all things considered. It avoids necessarily allotting culpability to serious akratic actions; these now may or may not be culpable. And yet it does not deny agency to anomalous beliefs and intentions. It says only that when an estranged attitude is present, the agency is not fully the agent's own because it cannot be fully and consciously self-determined. For this reason, it is to some degree psychologically unfree. From the third-person perspective, it may have only the same authority as does the subject's attribution of 'Jones believes that *p*'. There is nothing special about this third-person kind of authority. To think that it has first-person authority is to ignore a basic difference between attribution and avowal: that in avowal the subject makes up her own mind, in a justifiable way, what she believes, whereas in attribution neither of these is necessarily the case. Also that the subject's intention is different in each stance.

Anomalous self-ascriptions lack the subject's explicit, attention-focused awareness of an attitude that she *does have*, albeit inaccessibly at that time, where that attitude is not just accidentally overlooked in her thinking but is the

main reason for her self-ascription of the attitude she is consciously aware of having towards its object. Always present in an anomaly is an irrationality that needs explaining—a contradiction between what the subject self-ascribes as her belief or intention and what she relevantly but irrationally believes or does. Since first-person authority is about making up your own mind about something, what you decide to believe about it might well be false. As Moran points out, this first-person authority 'can be partial or hedged in various ways' (2001, p. 126). The perspective is that of first-person authority but the usefulness of the authority presupposed by this stance is often questionable and may sometimes be discounted.

It may be asked why we need Moran's (or anyone else's) account of self-knowledge to show that self-deception and akrasia require an estranged attitude for their explanation. Why not just say that people bury their unacceptable feelings, that this burying causes some mental phenomena to become unconscious and that this explains both the static and the dynamic paradoxes as the irrational consequences of this?

Such explanations remain detached from the most important problem that the anomalies present us with—the fact that they affect our lives because they affect our self-knowledge. This has practical consequences for our well-being. If we know, albeit unconsciously, what our real reason is for believing that *p*, this raises the possibility that we can discover what this reason is by bringing it to consciousness. Until we can do this, however, it may wreak untold havoc with our lives.

Moran's account enables us to understand the relation between the first-person authority of avowal, on the one hand, and the consequences of estrangement on the other hand, by spelling out the conceptual connections between the anomalies and a number of important concepts connected with self-knowledge—how we can know our beliefs immediately, how our avowals

of our immediately known beliefs can be substantive and how estrangement interferes with knowledge gained via transparency—not to mention the moral dimensions of the problem the anomalies present. It helps us to establish the way in which we exercise a very important kind of first-person privilege: a privilege that is agentive, inasmuch as it is based on how we question ourselves and the world, but essentially epistemic, inasmuch as it allows us to know ourselves. When we do not suffer from estrangement, we have a kind of epistemic access to our beliefs that no other agent could have. To me, the topic is an important one.

Chapter 1: Moran's account

The introduction has explained in general terms why I use Moran's account of self-knowledge in this thesis. In this chapter, I discuss his account in some detail, exploring more fully, in particular, how his major concepts are structured. I begin, in Section 1, with his basic claim that there is an intrinsic asymmetry between first- and third-person self-ascriptions of our beliefs and other intentional attitudes. In Section 2, I discuss rational agency, in Section 3, transparency and substantivity and in Section 4, estrangement.

Section 1: Asymmetry

To quote Moran on asymmetry:

> for a range of central cases, whatever knowledge of *oneself* may be, it is a very different thing from the knowledge of others, categorically different in kind and manner, different in consequences, and with its own distinguishing and constraining possibilities for success and failure. (2001, xxxi)

Moran's asymmetry claim is not only that there is a 'we and they' asymmetry between our knowledge of our own attitudes and our knowledge of the attitudes of others but also that this is mirrored in a *second* asymmetry, *within* our own self-knowledge, between two perspectives or stances towards an external object or situation that subjects can adopt in their intentional attitudes. These are the first-person and the third-person perspectives. We call a self-ascription from the first-person perspective an avowal and from the third-person perspective an attribution. His claim to this second asymmetry has been much criticised by deflationary objectors, who argue that immediacy is not really immediate at all but inferential, and that the first-person stance is reducible to the third-person stance.

In both stances, when a subject self-ascribes 'I believe that *p*' it is accepted that

she knows this immediately (without evidence or inference). So both stances are epistemic. However, the more fundamental of them, the first-person, deliberative perspective, is epistemic *plus*. When she is speaking from the first-person, avowable stance, the subject decides for herself what her attitude is towards some external object or situation by actively making up her own mind about this via her own reasoning capacity, so as to acquire *for herself* a belief, desire, intention or affective attitude towards this object or situation. For example, if someone asks her 'Do you believe that sea levels are rising?' she may reflect and then judge, on the basis of her own reasoning, that she believes this. In adopting this first-person perspective, she makes a commitment: she endorses, as true, her belief that sea levels are rising and she is prepared to defend it where necessary. This implies that the first-person perspective is normative. It is also *practical*; the subject *speaks for* it.

But having learned that avowals are based on *reasons*, one might object 'But so are attributions. So what is the difference?'

The difference lies in the fact that there are reasons that explain and reasons that justify. The subject who learned at school that Caesar crossed the Rubicon may be able to explain her belief but not justify it: 'I learned it at school' explains it without justifying it. Such a self-ascription is a third-person attribution, not an avowal. The justification required for an avowal comes from ourselves as agents who endorse and take responsibility for our reasons-based beliefs. We may have been taught how to justify our belief about Caesar at school, but to be able also to justify it now, years later, so that we can defend it, we must by now have 'made it our own' by being able to explain *why* Caesar crossed the Rubicon, so that we can produce this reason convincingly for someone else. Only when we can justify it can we successfully avow it. Otherwise, all we can do is attribute it to ourselves from the third-person stance and then *explain* why we hold it if asked. But explanation from the third-person

perspective does not have the first-person authority of an avowal. It tells us what caused us to believe that Caesar crossed the Rubicon but does not justify our belief. 'I learned it at school' says something about ourselves as subject—how we came to believe that Caesar crossed the Rubicon. Justification, however, requires information about the *object* of the belief, in this case, information *about Caesar*. For our self-ascription to be an avowal, we may be required to tell our listener why crossing the Rubicon made Caesar a great soldier. In this particular case we cannot do this and, indeed, are not trying to. We are quite happy with our attribution. Were someone to say 'Do you know why Caesar did that?' we must refer the questioner elsewhere. Certainly we are using our rationality in our self-ascription, but not in a justificatory way.

Moran uses Anscombe's[18] famous discussion of someone pumping water to explain the kind of 'why' we need in order to justify our first-person avowal: it is the kind that gives us the subject's aim in pumping water (Anscombe 1957, 14). What sort of reason could a person pumping water give that justifies, rather than only explaining why she is pumping water? Well, if she says 'These plants look too dry—they need water', that could justify it. If she says, 'I feel like watering something today', this explains her aim without justifying it.

So the two perspectives differ in their kind of authority. In both stances, the subject is assumed to have the authority of knowing *that* she believes that *p*. Even though she may not know why she has this belief, we accept that she does have it; we assume we can all know immediately, without evidence or inference, what our own attitudes are. But an avowal must also be justifiable: it must be solely about *the object*[19] of her belief, not about herself and not about anything that someone else has said about this object.

---

[18] This is discussed in Moran 2001, pp. 124-127.

[19] This is central to Moran's transparency condition, discussed in Section 3.

Consider the situation in which you self-ascribe 'I believe that *p*' for the first time. There are three different situations in which you can do this:

1) You can do it when you have just then made up your own mind, by deliberative reflection where necessary, that you believe that *p*. For example, you might say, 'I believe my mother died happy'. In this situation, you are avowing to others your belief about your mother.

2) You may self-ascribe your belief that *p* in a third-person, attributive way, either a) without wanting to avow it, although you could avow it if you wanted to, or b) without being able to avow it.

In 2a), you could avow your belief that your mother died happy (you can justify your reasons for believing this) but you do not wish to do so on this occasion because you know that the family will not believe your reasons. But in 2b), you have no idea why you think your mother died happy and you wish you could find out.

In 1), your listeners are right to assume that your self-ascription is an avowal. This is because they assume that you believe that *p* for a reason, that this reason is about the object of your self-ascription and that you could tell them that reason if you were asked. In this case, the reason must be about your mother. It could be that you and your mother were close and that you could tell from the way she looked, spoke and acted, that she was happy. This accords with Moran's claim that avowals are based on reasoning that can justify the avowal. The subject can defend his beliefs because he holds them for reasons about the object of the self-ascription that he can produce if necessary.

In 2), however, you may or may not be able to avow your belief that *p*.

In 2a) you can do so but do not wish to.

In 2b), however, you can give no reason for your belief, and if someone were

to say, 'You weren't even there at the time. You have no reason for saying that', you can give no answer. So in 2b, the subject's self-ascription does not imply avowability. Thus, on Moran's account, an attribution does not necessarily imply the first-person authority of an avowal that the subject can justify.

In spite of this, however, we often *assume* avowability whenever someone says 'I believe that *p*', unless the context clarifies this. It is *normal* to suppose that the subject holds her belief for a reason of her own. When we find out that she is not prepared to argue for it when challenged, we may realise that she has attributed it to herself instead of having avowed it. Moran argues that in this 2a) or 2b) kind of situation, the speaker has told us only something about herself; she has told us what is *in her mind*. She has *attributed* this belief to herself. Her attribution is empirical only, not also normative. She tells us no more than an *empirical fact* about herself.

But because her listeners may assume avowability in 2a) and 2b), this makes it seem as if all self-ascriptions are the same. However, there are a great many beliefs that we hold in the 2b) way because we have learned them and never questioned them. Some of them we might try to defend, because we think they must be correct. Some of them we change when society changes its mind about them. Some of us might change our minds about this because we become convinced by new reasons for doing so, while others might just 'go along' with this new idea. They realise that they should have some good reason for being convinced, but they do not try to find one.

So not all situations in which we attribute some belief to ourselves are also avowals. Some are, some are not. This means that the assumption we usually give to attributive self-ascriptions (that they are avowable) is not always correct. It may, however, play a role in explaining the deflationary assumption that first- and third-person authority are symmetrical rather than asymmetrical. In fact, third-person authority is not especially authoritative at all. It does have

the epistemic authority of agency—a belief attributed to oneself is usually thought to be known immediately, whereas when I say 'Jones believes that *p* but goodness knows why' when I have just overheard someone announcing Jones' belief, I cannot justify my claim because I have no idea whether the person I have overheard is right or wrong. Justification, as against explanation, as well as being about the object of the belief, must also be successful from the first-person perspective. The belief must be transparent to its truth[20] in the eyes of the self-ascriber, although not necessarily in the opinion of her listeners. It is up to the self-ascriber what she believes.

Moran tells us that

> What we're calling a theoretical question about oneself, then, is one that is answered by discovery of the fact of which one was ignorant, whereas a practical or deliberative question is answered by a decision or commitment of some sort, and it is not a response to ignorance of some antecedent fact about oneself. (2001, 58)

It is the fact that the subject has made up her own mind about it and can justify it, together with its practical application in her life, that explains the special authority of the first-person.

There is a further complication. When estrangement is present, the subject's avowal *necessarily* fails. [21] First-person authority does not encompass estrangement of any kind. This means that in an avowal that fails because of estrangement, there is necessarily a passive element in the subject's reasoning because of the unconscious presence of the estranged attitude. This implies

---

[20] If I believe that *p* I must also believe that *p* is true.

[21] When there is no estrangement, the avowal may fail for empirical reasons, such as fatigue. However, such reasons imply that the attitude being avowed is not settled. Moran's account does not encompass unsettled avowals.

that the avowal is not fully self-determined. Because the subject suffers a degree of passivity, her thinking about it is psychologically unfree.

So there are a number of things to consider in unravelling the differences beween attributions and avowals. One of these is whether cases of 'knowing how', as against 'knowing that', have the first-person authority of an avowal rather than just the third-person authority of an attribution. Know how is a kind of *practical knowledge.* Agents sometimes have knowledge how to do something – how to ride a bike, how to kick a football – apparently without propositional knowledge concerning these skills. They therefore seem unable to avow their know how. Let us explore this question by using examples. By comparing two of these, the farmer who can predict rain and the chicken sexer, I argue that we can see why Moran does not include cases of 'knowing how' in his account of the asymmetry between attributions and avowals.

Example 1: The farmer who can predict rain

This farmer cannot say why the clouds look as they do just before it rains. But as well as the evidence of his own eyes, he has the evidence of his family who have consistently been right that it is going to rain whenever the clouds look as they do just before it rains. This example, it could be said, is a case of *knowing how*—how to predict rain. Let us suppose the farmer says, 'I know how to predict rain'. This self-ascription is attributive. It is said after the event (after he has learned how to predict rain) and he is just telling us this empirical, psychological fact about himself, viz. that he knows how to predict rain.

Could his self-ascription also be avowable? Suppose that having looked at the clouds the next day, the farmer shouts, '(I believe)[22] it's going to rain!' It would seem that this latter self-ascription can be an avowal. The farmer has assessed

[22] The 'I believe' may be omitted but understood as implied.

the situation and made up his mind that it is going to rain. If challenged, he can point to the way the clouds look and he can put this into words to some extent. This implies that when, earlier, he says, 'I know how to predict rain', this self-ascription also has first-person, deliberative authority. He has an immediately known, practical, justificatory aim in avowing his belief—he must quickly (say) shelter his vegetable patch or put his cows in a different paddock.

But not all cases of 'knowing how' are like this, as our next example shows.

Example 2: The chicken sexer

As I understand it, there was a time when some successful chicken sexers did not know why they could sex chickens correctly. They knew only that they could do so. Later, someone discovered that male chicks straighten their legs when picked up while female chicks draw their legs up under them. Some chicken sexers who did not know this fact were apparently unconsciously observing it and applying this unconscious observation in deciding the sex of each chick. (Or they were keeping their knowledge secret from their customers! That seems far more likely to me, but let us assume for the purpose of this discussion that this chicken sexer was not doing this.)

Suppose that the chicken sexer says, 'I am an excellent chicken sexer'. Is his self-ascription an avowal or an attribution? Well, although he does not know why he is such a good chicken sexer, he does know why he believes he is. Like the farmer who learned about clouds from his relatives, he has proved, over and over again, that the chickens he says are male always grow up to be roosters while the chickens he says are female always grow up to be hens. So although he cannot say why this happens, his belief that he is good at chicken sexing seems to be rational. He knows that he can prove his expertise, if asked, on the backed-up evidence of many grateful chicken farmers.

There is, however, an important difference between the farmer who can predict

rain and the chicken sexer. While the former can avow *this is the way to predict rain* – by looking at the very things that produce rain—the clouds—and by pointing out (say) their colour or size – the chicken sexer does not know that it is by looking at the chicks' legs that he succeeds in his task. He cannot provide an answer to the question how does one ϕ by the demonstrative *this is how one φs*, and therefore cannot satisfy the propositional component of knowledge how (Stanley 2011). He may know this fact unconsciously, but because it is unconscious he cannot rationally link it with the chick's sex, whereas the farmer who predicts rain can rationally link the clouds causally with impending rain. Surely we must say that the chicken sexer's self-ascription fails to achieve much in the way of authority. When he says, 'I believe that this chick is male', he may be in deep trouble if someone asks him *how* he knows this. I suggest that when his many chicken farmer customers say 'He is an excellent chicken sexer', *they assume that he knows how* he sexes chickens. If his listeners discover that in fact he does not know how he does it, they might find another chicken sexer.

On my view, because he lacks the capacity to answer the question "how do you ϕ," the chicken sexer's self-ascription is an attribution, not an avowal. However, this is a difficult case to fit confidently into either the first- or third-person stance, and it might be one reason why Moran does not include cases of knowing how in his account.[23]

The above discussion has argued in support of Moran's claim that there is an asymmetry between first- and third-person self-ascriptions of one's intentional

[23] The chicken sexer case raises for us the question of the relation between rationality and consciousness. The chicken sexer cannot justify his claim because he cannot consciously justify it. In Chapter 6, I also argue this way against Tappolet's claim (Stroud and Tappolet, 2007, 97-120) that in Arpaly's case of Emily (2000, 504) the subject (Emily) can act rationally on an unconscious attitude.

attitude. Next, I briefly introduce Moran's approaches to rational agency, transparency and substantivity.

The rational agency dimension of the account

Active, rational agency is basic in Moran's account. His account of such agency differs from those of other philosophers[24] not only in adopting the first-person perspective on active agency but also in stressing the importance of a certain kind of activity. He does this by drawing on Kant's distinction between, on the one hand, pure apperception, signified by the 'I think' and involving the active exercise of agency and on the other hand, empirical apperception or inner sense, where the agency involved is more passive, as in feeling a sensation[25]. In Section 2 of this chapter, using Boyle's article (2009, pp. 133–164), I briefly describe these two kinds of self-knowledge, explaining why the kind gained by active agency, using the deliberative stance, is more fundamental than the kind gained by the more passive attributive stance, why this latter kind of activity has only third-person authority and why this is central to Moran's position.

The transparency dimension of the account

On Moran's account, when a person asks herself 'Do I believe that *p*?' she ordinarily treats this as equivalent to the question 'Do I believe that *p* is true?' even though the question 'Do I believe that *p*?' is a question about herself while the question 'Is *p* true?' is a question about the external world. Empirical transparency accounts of self-knowledge are confined to explaining that we form the belief that *p* by forming the belief that *p* is true. For example, Dretske

---

[24] Gallois (1996) and Shoemaker (1994) claim that rationality is essential to self-knowledge but do not consider the first-person perspective in their accounts.

[25] Of course, even to mention a sensation one must be able to conceptualise the feeling, as being one of pain or itching or hunger or even just a sensation. But conceptualising it is different from exercising one's active agency with the aim of making up one's mind about some matter that involves it.

(1994) uses transparency to argue that we know our own minds only by inference, from perceiving the external world. Fernandez (2003, 352–372) argues that believing that one believes the propositions that are supported by evidence is a generally reliable belief-forming process. However, since both of these accounts are deflationary, they cannot consider the subject's conscious, immediate, first-person awareness of her mental states in their explanations in order to discover whether the beliefs we form have first- or only third-person authority. Moran uses transparency from a first-person perspective, to explain how subjects can acquire avowable self-knowledge (know their own attitudes by creating them and being able to justify them) and thus how we can decide whether a self-ascription has first- or third-person authority. In his account, active first-person authority and estranged attitudes are structured as opposites. This structure gives us the link between estrangement from self-knowledge on the one hand and the anomalies of self-deception and akrasia on the other hand by explaining how these anomalies occur. It argues that transparency accounts are normative and so subject to active, rational agency, thereby linking transparency both to successful rational agency and to its anomalies as being the opposite of this. I set out his transparency account in Section 3.

The substantivity of psychological states

My final discussion of asymmetry between first- and third-person 'I' self-ascriptions concerns the substantivity of both the attitudes that can be self-scribed—attributions and avowals: how our mental states can have their own independent, epistemic, psychological status that we can know immediately (without inference or evidence). This affects the asymmetry because if our mental states are not substantive, they cannot be epistemic either, and thus the debate about asymmetry loses its significance. The authority of the first-person avowal would no longer be special: our knowledge of our beliefs about

ourselves would be evidence-based.

The long-standing influence of the Cartesian picture of introspection created a skepticism about whether our awareness of our own minds can be of any cognitive significance. While breaking with the Cartesian claim that introspection is a kind of perception, Moran sees first-person awareness of our mental states as being of something that is substantial, representing a genuine cognitive achievement rather than being either 'the shadow cast by certain features of our linguistic practices' or of something that is not cognitive or epistemic (2001, p. 21-23). Wright (1986, 401), for example, has argued that our first-person judgments of our mental states are extension-determining rather than extension-reflecting. By this he means that a person's best opinion about his intention, for example, determines or constitutes it rather than, having already constituted it, tracking it later in the way one tracks something that exists independently of that person's opinion as to what it is (Moran 2001, pp. 22 to 23). If Wright is correct, our thoughts, desires and feelings are not substantive. Once we have constituted them (when we do), they do not exist in their own right, independently of our thinking or feeling them. Moran's claim that, on the contrary, our mental states do have their own substantial, epistemic status, is congruent with his emphasis on common-sense psychology as well as with the rational agency dimension of his account and its links with transparency.

Moran adopts a commonsense realism towards our mental states. This realism is grounded in his claim that first-person authority is founded on rationality and transparency, and supports his claim that self-knowledge is substantive. He claims that attitudinal self-knowledge represents a genuine cognitive achievement involving the awareness of some independently obtaining psychological state(s) and as having truth-conditions that are independent of the making of the judgment (so that a first-person belief may be false, for

example).

Moran argues that a first-person awareness of some belief of our own is an awareness of something that is substantive. Self-knowledge is 'substantive' for Moran inasmuch as it is epistemic, representing a genuine cognitive achievement involving the awareness of some independently obtaining state of affairs and having truth-conditions that are independent of the subject's making of the judgment. It represents a genuine cognitive achievement (2001, p.3) rather than being trivially or conventionally or constitutively true, for example. By this he means that psychological states such as beliefs can be substantive without being either inferential or seen by the 'inner eye' and thus (on a Cartesian approach) incapable of error:

> I wish to defend a view of first-person awareness that sees it as … substantial … but which nonetheless breaks decisively with the Cartesian and empiricist legacy.

An agent's reflection on whether to believe that *p* makes use of both rationality and transparency. Reflection occurs via her reasons-based consideration of evidence about how the world is. Thus it provides *reasons* for her to use in making up her mind whether to believe that *p*. She can know these reasons *immediately* because she has formed them for herself.[26] They are substantive because, via transparency, they are grounded in how the world is. No more than minimal rationality is required; we exercise our reasoning capacities sufficiently by being able to form intentional states even when these are only minimally keyed to objective features of the world.

The question of substantivity cannot be divorced from that of immediacy in Moran's account. The basic concept of immediacy that Moran claims is not about anything incapable of error but about an awareness that is not inferred

[26] The question of *how* we know our beliefs immediately is discussed in Chapter 4.

from anything more basic (p. 11). He points out that we can usually know our own bodily position without having to observe anything, internally or externally; we can also sometimes guess what time it is without observation or evidence of any kind (p. 19). Immediate self-knowledge is also substantive in that the subject can immediately detect an independently obtaining, psychological state of affairs (p. 13); such discoveries are cognitive achievements.

Martin (1998, 107 to 108), Boghossian (2003) and Macdonald (1998), among others, have given reasons for thinking that self-knowledge is substantive. Martin discusses cases where the subject has a strong conviction that *p* is true but lacks sufficient evidence to be willing to commit to holding the belief that *p*. These can include cases of mild self-deception, such as the proud father's conviction that his son is a fine painter in spite of lack of evidence for this (114–115). This father is motivated to slant the evidence in favour of his son.

Let us suppose, Martin argues, that the father sincerely self-ascribes, 'I believe that my son is a fine painter'. The belief that he self-ascribes, although he is sincere, is false; this father actually does *not believe* that his son is a fine painter. Basically, he *wishes* that his son were a fine painter, but is estranged from this wish. Thus there is no constitutive relation between his apparent second-order belief, 'I believe that my son is a fine painter' and his first-order belief, 'My son is a fine painter', because the first-order belief 'My son is a fine painter' is false and its paired, apparent second-order belief, 'I believe that my son is a fine painter', does not exist. What exists is his estranged, first-order wish that his son be a fine painter. But if one can self-ascribe a belief that one does not in fact have, the second-order belief that is normally paired with this first-order belief must be distinct from its paired first-order belief because in some cases, the second-order belief does not exist.

Thus, Martin argues, some belief pairs of the same fundamental type must be

distinct existences. This must be the case even if it is also true that the first-order belief constitutes the second-order belief in providing the content for the second-order belief. First-person authority, Martin claims, is not a matter of a belief's mode of existence but of how the subject has arrived at it. If the second-order belief is false because there is in fact no first-order belief pairing it, first-person authority will have failed.

Boghossian (2003) also gives us an argument in favour of substantivism. He first points out that to say that self-knowledge is based on nothing really means to say that it is based on nothing empirical, and, in this way, that it is cognitively non-substantive (p. 76). But this, he says, cannot be right. He gives several reasons for this, the most important of which, he suggests, is that a substantive construal of self-knowledge presupposes that self-knowledge is both fallible and incomplete. But if the subject can misconstrue her knowledge of her attitude, this can be explained only via the idea that she is not in a favourable *epistemic* position with respect to this knowledge. If self-knowledge is *non*-substantive, we cannot explain its admitted shortcomings, since only epistemically substantive self-knowledge can be either right or wrong.

Boghossian says:

> the difference between getting it right and failing to do so (either through ignorance or through error) is the difference between being in an epistemically favorable position with respect to the subject matter in question—being in a position to garner the available evidence—and not. To put this point another way, it is only if we understand self-knowledge to be a cognitive achievement that we have any prospect of explaining its admitted shortcomings. (2003, p. 76)

Macdonald (1998, p. 138, n. 16) gives us another reason that supports substantivism. She points out that one can ask oneself, 'Do I believe that *p*?' and proceed to reflect on this question, and that in this situation there is a first-order propositional state, *p*, which is being reflected on and which plays the

role here of being the theoretical–descriptive base for possible psychological self-ascriptions. Thus it must be distinct from the second-order state *via which* the subject is reflecting on it and can arrive at different conclusions about it from a prescriptive[27] point of view.

The asymmetry that Moran points to between our knowledge of our own attitudes and our knowledge of the attitudes of others is not denied by other writers; it is fully accepted. His claim that first-person self-knowledge has a special and unique authority also has much support, although its immediacy, which is what makes its authority special, has been criticised. However, his claim that there is a second asymmetry, between a person's third-person and first-person knowledge of her own intentional attitudes, has met with a volume of criticism.

I have tried to show above that this second asymmetry is intrinsic and thus irreducible. It is based on the fact that the subject may have either of two intentions in self-ascribing her belief (or other attitude): i) simply to inform the listener of her belief by stating it or ii) to avow it to her listener by ascribing it to herself in order to defend it if necessary. Both stances are reasons-based and thus agential, although they may demonstrate a degree of failure of active agency. Both are also epistemic and substantive. But the attributive stance states only an empirical, reason(s)-based fact about the subject's mind, while the avowable stance uses reasons in a normative and practical way, involving the subject's endorsement of the truth of the belief and her commitment to it. When avowal fails, the avower's intention is thwarted. In Chapters 5 and 6, I argue that in self-deception and akrasia the subject's estranged motive is to

[27] Here Macdonald is using Moran's 1994 distinction between theoretical–descriptive and prescriptive bases for psychological ascriptions, a forerunner of his later theoretical/deliberative distinction. She introduces this distinction in Macdonald 1998, p. 124.

avoid confronting an unacceptable fact about herself.

In Section 2, I set out Moran's account of active rational agency, pivotal to his explanation of first-person authority.

Section 2: Active rational agency

In this section I will set out the two forms of activity that constitute, on Moran's account, two different kinds of self-knowledge: one gained by active agency and the other by more passive agency. These correspond to the two stances or perspectives, the deliberative and the attributive, that he postulates, and to their two kinds of authority, the first-person authority of active agency and the third-person authority of more passive agency, such as our knowledge of our sensations. I will also set out Boyle's (2009) argument that we can represent ourselves in two ways: actively, by making up our minds about *p*, and more passively, by expressing our knowledge that we are in pain. These two ways of representing our self-knowledge are asymmetric. But the first, actively making up our minds, is primary, because without it the second would not be self-knowledge at all, but only parroting.

Moran's account of active, rational agency and its relation to first-person authority is very different from orthodox, contemporary accounts of the place of rationality in an account of self-knowledge. Shoemaker (1994, pp. 249–314), for example, is a major advocate of an orthodox rationality theory. Shoemaker's deflationary, functionalist account of self-knowledge uses the idea that no rational person who has acquired the concepts of mental states generally, including the concepts of belief, desire, intention, pain, happiness and other similar mental states, could be self-blind and thus incapable of self-knowledge: that is, he claims that self-knowledge is an essential part of our rational nature. This is consistent with Moran's approach. But Shoemaker also claims that no inner (mental) process could enable one's knowledge of one's own intentional

attitude to differ from its apparent behavioural manifestations. 'Available beliefs', he claims, rather than phenomenologically conscious mental states, are normally sufficient for self-knowledge (2009, p. 31). Thus self-knowledge, on Shoemaker's theory, is justifiable only in epistemically externalist terms, leaving out any need for the subject's own first-person, conscious awareness of her attitude and thus attempting to remove the asymmetry between the first-person perspective on self-knowledge and the third-person perspective.

Moran's rational agency account is thus unorthodox in two ways. Firstly, unlike Shoemaker's account, for example, it insists that the first-person perspective is asymmetric with the third. This is related to the second unorthodoxy: his account harks back to Kant's distinction between pure apperception, signified by the capacity to say 'I think that *X*', which involves the active exercise of agency, and empirical apperception or inner sense, where the agency involved is more passive, as in feeling a sensation. Kant argues that the passive kind of self-knowledge depends on the active kind, in that we would not be able to know our own sensations at all without any capacity for the active kind, since it is this latter that makes us 'knowing beings'.[28] Moran uses this Kantian idea to claim that that capacity to say 'I think that *X*' is what makes rational agency fundamental to self-knowledge.

Boyle: the uniformity assumption is wrong

The claim that there are two kinds of self-knowledge, of which the active kind is fundamental while the passive kind depends on it, implies that immediate, authoritative self-knowledge cannot all be explained in just one basic way. Sensations, although known passively, can clearly be known with immediacy and authority. Does this mean that all our self-knowledge can use sensations

[28] See Kant 1929, B67–8, A107, B132, B153 & B278.

as their exemplar? Matthew Boyle (2009, pp. 133–164) calls this idea that there must be just one basic way of knowing all our mental states 'the uniformity assumption' and argues that it is wrong: we can use Kant's active/passive distinction to differentiate between an active kind of self-knowledge that has *first-person* authority and a passive kind that has only *third-person* authority (p. 134) and that it is our capacity to make up our own minds that explains our capacity to say 'I think that *X*'.

Briefly, Boyle argues that to account for self-knowledge of any kind we must distinguish between behaviour that merely demonstrates that we are in a certain mental state and behaviour 'that expresses a representation of oneself as in a state of the relevant sort' (p. 135): that is, the capacity for self-knowledge involves the capacity for self-representation. The relevant sort of mental state is thus one in which the subject understands what she is saying in expressing her self-knowledge, as against merely 'parroting' this self-knowledge. This obvious condition implies something more basic, however: a self-knower must represent her own condition as being of a certain kind. What is this kind? Boyle's argument is that the speaker's 'implicit grasp that he has the power to make up his mind is a condition of his understanding the first person at all' (p. 155) and that 'the power to represent one's own deliberated attitudes presupposes the power to know one's own deliberated attitudes in the way that Moran specifies' (p. 147): that is, only an account of self-knowledge that recognises the distinctness and fundamentality of the kind of self-knowledge that Moran (2001) has already identified as deliberative avowability (based on making up one's mind as a rational agent) can account for the relevant sort of representation (Boyle 2009, p. 143). The claim is that to be a *self*-knower at all, a person must be able to use the term 'I'. To have *this* capacity is to have an implicit grasp of the fact that he has the power to make up his mind to believe that *p* or to do *X*: that is, to be an active, rational agent.

Boyle's argument for this claim begins with the difference between a human speaker who says 'I am in pain' and this same sentence uttered by a parrot. A parrot can be trained to cry out 'I am in pain' just whenever it is in pain, but the parrot does not understand what it is saying. Its behaviour manifests pain but not the parrot's knowledge that it is in pain. The human speaker who avows, 'I am in pain', though, knows that he is. His avowal is not an automatic response that enables him or someone else to suppose that he is in pain. If our avowals can express knowledge of our own minds, we must distinguish between the sense in which the parrot expresses that he is in pain and that in which a competent human speaker expresses that he is in pain. Boyle calls the first (the parrot's) a manifestation sense, or expression$_M$ and the competent speaker's a representation sense, or expression$_R$. We can talk about someone expressing self-*knowledge* only when he is saying 'I am in pain' in the expression$_R$ sense because only in this sense is he representing his own state as being of a certain kind. To say 'I am in pain' in the expression$_M$ sense, or to cry out in pain in the expression$_M$ sense, 'would so far be exhibiting no more self-knowledge than is exhibited by our imagined parrot' (p. 145).

Boyle then argues that any theory of self-knowledge must account not only for our having mental states but also for our representing (expressing$_R$) our own mental states as our own (p. 146). This raises for Boyle the question of whether this accounting for 'requires crediting the subject with a special kind of knowledge of his own deliberated attitudes'. Boyle's answer is that it does. As he points out, expressivist writers and those who suggest that we have 'monitoring mechanisms' both have to explain the relation of expressing$_M$ their sensation of pain in a way that allows that the speaker knows that they have this pain and that it is their own pain. Unless their avowal can *represent* (express$_R$) their pain they cannot be credited with *knowing* that they have it and that it is their own, any more than when the level of mercury in a thermometer monitors expresses$_M$ the temperature we can say that the

thermometer knows what the temperature is.

A person's being able to recognise that the pain she avows is her own pain implies that her avowal is from the first-person stance. This is important for being able to explain the immediacy of first-person self-knowledge. Only an account of self-knowledge that can explain how it is that our avowals have a special form of authority and that we have a special, irreducible access to them can be successful.

So what sorts of ability must the competent avower of 'I am in pain' have that the parrot does not have? Boyle argues (2009, p. 150) that a creature who can express$_R$ her pain can also draw conclusions about her own beliefs in the way that Moran describes in his account. She can recognise relationships among the concepts in her sentences, can grasp the relation of their content to the contents of a system of possible other claims and can reflect on her grounds for holding a given claim to be true. In other words, she can deliberate about why she believes that *p*, and can make up her mind about whether *p* in a way that conforms to Moran's transparency condition. When she sincerely affirms that *p* she affirms something she takes to be true (p. 151). Boyle concludes that the kind of self-knowledge that Moran calls his deliberative or avowable stance is fundamental.

It is also first-personal, in that the expression 'I' refers to the speaker. 'I' is self-referential. However, one can understand how to use the English word 'I' without understanding its significance. To understand its significance one needs to know its links not only with knowing that one is in pain but also with knowing that *p* because one has determined that *p* for oneself—because one has made up one's own mind whether *p*. It is because one can make up one's own mind about whether *p* that one can also know that one is in pain. But knowing that one is in pain is a different kind of self-knowledge, a more passive kind, than being able to avow that *p*. What is the same in both cases, Boyle

suggests, is that knowing one's own mind is a precondition of self-consciousness. Only a creature who can answer the appropriate kind of 'why?' questions (questions about reasons) can avow her beliefs, but creatures who can express$_R$ their pain and joy must also *be able to* make up their own minds whether *p*. The active kind of self-knowledge is therefore primary.

Boyle concludes that the difference between the active and the passive forms of self-knowledge is a difference between two kinds of agency. If he is right, the uniformity assumption is false and our knowledge of our own beliefs, desires, intentions and affective attitudes, when it conforms to Moran's transparency condition and is minimally rational, is asymmetrically first-personal.

In Section 3, I discuss Moran's two perspectives with respect to transparency.

Section 3: Transparency

The concept of transparency denotes an empirical fact first suggested by (among others) Edgeley (1969, p. 90) and Evans (1982, p. 225), who pointed out that we can gain self-knowledge by looking not inwards into our minds but outwards to the world. When asked, 'Do you believe that sea levels are rising?' we typically answer not by looking inwards to see whether our minds contain a belief about sea levels but by considering external evidence about whether sea levels are rising. In making up our minds that the proposition 'Sea levels are rising' is (say) true, we judge and then, after reflection, may form the belief that sea levels are rising. In these cases our belief that *p* is said to be transparent to the truth of *p*.

Most transparency accounts are perceptual, differing from Cartesian and post-Cartesian accounts (where the subject looks inwards, into her mind, to discover what she believes) mainly in that the subject looks outwards to the world to

discover what she believes or intends to do. But the main difference between empirical transparency accounts and Moran's account is that on Moran's account the subject does not *discover* what she believes or intends to do by looking somewhere in the world where this answer already is, but rather makes up her own mind about this: that is, *creates* the answer for herself. The facts she uses as evidence for doing this are already there to be reflected on, but what she does with this evidence, in reflecting on it, is *up to her*. She can do it well or badly, successfully or unsuccessfully. Prima facie, it will have first-person authority if it is minimally rational and conforms to Moran's transparency condition.

'Edgeley–Evans type' questions are normally treated as practical. The subject answers them by exercising her capacity as a rational agent in considering the facts that constitute evidence for or against them, to reach her conclusion. Thus her deliberations are normative, producing a judgment about what she ought to believe or do. However, Moran argues that she does not stop at that point, she decides for herself what she actually *does believe* or *will do* (2001 p. 59). She makes up her own mind about this, sometimes even when it conflicts with what she thinks she ought to believe or do. The conflict is between (say) the intention to *a* and the desire to ~*a*. These may come apart after she has formed her intention. She may end up deciding not to carry out this intention because of this conflict. Indeed, when she intends to do what she ought to do without making up her own mind about it first, but rather because it is what she has been taught to do, her knowledge of this intention will have only third-personal, empirical authority. She may end up not carrying out this intention, either.

Transparency accounts of any kind seem to introduce a problem that does not beset neo-Cartesian accounts of self-knowledge. This problem is that our belief is something about us, whereas *what* we believe is something about the world.

How can we form a belief which is something about ourselves (what *we believe* about something) by considering not ourselves at all but something in the world? Why don't we consider evidence about whether we believe that *p* by considering our own behaviour, for example? How can anything about the world tell us what *we believe* about it?

Evans' famous quotation seems to give us the answer:

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

Evans' point is that I look in the same place to find my *belief* about another war as to think about *about* another war. To form a belief about another war I consider the world. Having considered it, and having, on the basis of this considering, formed my belief that (say) there will not be a third world war, then *when I am asked this question* and sincerely reply 'No, I believe there will not be a third world war', my self-ascription tells us that I know that I have this belief.[29]

But empirical transparency accounts do not accord with our intuitions about the need for the self-ascriber to be able to give reasons for her self-ascription when it is appropriate and reasonable for us to ask her to do this. Transparency might be able to show that the self-ascriber knows *that* she believes that *p* but

---

[29] In itself this does not seem to justify my self-ascriptions in other situations, where I have not just been asked and answered a question such as Evans presents. Nor does it seem to explain how I can know my already existing beliefs. These issues will be discussed in Chapter 3.

cannot in itself show that she knows *why* she believes that *p* at times when it seems important that she be able to do so.[30] However, when her thinking is muddled owing to (say) fatigue and her self-ascription is not settled[31] its first-person authority may fail at the time if the belief she forms is false or unintelligible or the intention less than minimally rational. Still, she speaks (albeit unsuccessfully) from the first-person perspective. From the third-person perspective of an observer, however, her 'avowal' has third-person authority only. Or, if she is deluded, it has none at all. Unsettled attitudes due to fatigue or similar temporary impediments are irrelevant to Moran's account of the relation between self-knowledge and the anomalies because the subject's lack of mental clarity at those times is not due to estrangement or alienation in any philosophically interesting sense.

The role of avowal in transparency

Moran calls a first-person, present-tense, deliberative, attitudinal self-ascription an avowal, and explains 'avowal' as follows:

> 'Avowal' is defined as a way of answering a question about one's belief or other attitude that obeys the 'Transparency Condition', hence a form of self-knowledge that is immediate because transparent to a corresponding question that is directed outward, upon the world[34] (2004, p. 424).

On Moran's account, when the subject can answer 'Is *p* true?' in the same way as she answers 'Do I believe that *p*?' *and can avow* this answer, her avowed belief can be a form of immediate self-knowledge because her answer to 'Do

---

[30] Fernandez (2003) offers an empirical account of transparency by arguing i) that evidence that *p* will lead to a belief that *p* and ii) that believing propositions supported by evidence is a reliable belief-forming process. But a reliable belief-forming process does not differentiate between knowledge of our beliefs that has first-person authority and knowledge of our beliefs that does not. The subject's ability to say why she believes that *p* when this is appropriate is crucial to its first-person authority.

[31] Moran's account deals only with settled attitudes.

you believe that *p*?' is transparent to her answer to 'Is *p* true?' rather than being based on evidence about herself as believer. This gives us a simple way of understanding immediacy: our knowledge that we believe that *p* is immediate when our belief that *p* is transparent to its truth as we see it.

The components of Moran's transparency condition

For an avowal of a belief to have first-person authority, Moran's successful avower must conform to the following:

- She must reflect on whether *p* (where reflection is necessary) from the deliberative stance.
- She must focus her deliberations solely on the external object of the belief, not on facts about herself as believer. The reason(s) she can give for believing that *p* must make sense to her solely as reasons pertaining to the object of that belief. They must not have been determined by motivating forces independent of those reasons.

The first of these components simply states that for transparency to obtain at all, the subject, in her deliberations, must be using the deliberative perspective in reflecting on whether *p*. This means she must treat the two relevant questions, 'Do you believe that *p*?' and 'Is *p* true?' as if they were the same question, since, on Moran's account, the question of transparency does not arise in the empirical, theoretical stance, where the subject is neither considering the external object of *p* nor endorsing the truth of *p* in her self-ascription. Of course, they are not the same question from a third-person perspective; I can believe that *p* and also believe that *p* is true—both by using transparency—while *p* is in fact false.

The second component brings us to Moran's view on estrangement. When there is an estranged attitude unconsciously distorting her reasoning, producing a false or irrational belief, the subject necessarily *cannot* conform to

the transparency condition because inaccessible facts about herself as believer bias and constrain her reasoning about the belief's object.

This second component also shows us why 'I believe that Caesar crossed the Rubicon because I learned it from an excellent history teacher' does not conform to transparency. It is because where the subject learned this fact about Caesar is not about Caesar. It is a fact about the believer, not about Caesar. Such an 'avowal' therefore fails. A fact about Caesar would be if the believer could say, 'Caesar crossed the Rubicon because he figured it would give him a great victory, so enhancing his reputation as a great soldier'. Even if this claim is incorrect, the avowal might still count as an avowal, imperfect though the believer's knowledge of Caesar has been shown to be. I continue this discussion shortly.

But first, let us continue to discuss Moran's claim that the explanation a subject can give of his avowal must conform to transparency. I turn now to the example of Abigail, who eats too many chocolates, to do this.

Abigail 'avows' that she envies her friend because the friend is better than she is at colour coordinating her wardrobe. At first glance, this 'avowal' seems to meet Moran's transparency condition and so seems to have first-person authority. The friend *is* very good at colour coordinating her wardrobe; it is a reason for envying her friend that is about the friend rather than about Abigail herself; it makes sense to Abigail and to others and, as far as she knows, Abigail *is* focusing her deliberations on her friend rather than on anything about herself. So this reason *seems* to fit components i) and ii) of Moran's transparency condition.

But her friend's creativity with colours is not Abigail's real reason for envying her; she is estranged from her real reason because she has repressed it. Unknown to Abigail's conscious awareness, her admiration of her friend's

creativity with colours is masking her real reason for feeling envious of her. The real reason for her envy is her estranged belief that she (Abigail) is not popular among her friends because she is obese. This belief is the real reason for her envy of her friend. Her conscious belief, that she envies her friend's creativity with colours, functions as a substitute for (a rationalisation of) her estranged real reason, enabling her to avoid becoming aware of the latter. She has failed to conform to component ii) of Moran's transparency condition because her reason for envying her friend is a reason that is about herself, not about her friend. Therefore, her avowal is unsuccessful. It is not solely about its object, her friend, after all; it is largely about herself. It is a pragmatic or 'means to an end' self-ascription. She knows *that* she envies her friend, but she does not know her real reason for doing so. Her avowal allows Abigail to praise her friend for her colour coordinating, thereby expressing envy in a socially safe way: it is acceptable to express envy openly as a way of praising a friend. To anticipate Chapter 5, I suggest that this case might be an example of self-deception. Even when the real reason for her envy is consciously held, Abigail may 'block off from it' at times in order to achieve a self-ascription that might avoid it but that does not really do so. Note that given that her real reason is inaccessible to her, her reasoning is otherwise faultless: her partially active, agential rationality is functioning as well as it can in the circumstances. But where the real reason for her envy of her friend is concerned it is passive, not active. She has no *immediate* knowledge of why she envies her friend, and thus from the third-person point of view she is merely a third-person spectator of her belief. This example shows that an avowal can have a great deal about it that is rational, and yet be an example of a self-ascription that *seems* to be an avowal but that has failed to conform to transparency because it suffers a degree of psychological unfreedom via having an estranged attitude.

We also need to ask about the status of her estranged belief, 'I envy my friend because I am obese and unpopular while she is popular because she is not

obese'. Because this latter belief is estranged, Abigail knows it only in a tacit way[32]. Although she knows that she envies her friend she does not know why.

The transparency condition is Moran's argument for his claim that a successful avowal can deliver immediate, epistemically substantive knowledge of the subject's belief that has first-person authority. The subject's knowledge of it can be immediate (groundless) because it is her own belief, formed in a way that is 'up to her'. Thus she can know it *for herself*. It is, as Moran says, *her business* (2001, p. 123), as a rational agent, whether she believes that *p*. It has justifiable first-person authority because her business 'counts for something', counts as determining her belief. Her knowledge of it is also substantive, Moran claims, because the exercising of her own reasoning about the attitude's external object actually determines what her belief *is*. She can know immediately what her belief is because she has formed it for herself, based on external evidence that she herself has examined. It is wholeheartedly *hers*.

Avowal, which Moran describes (2001, p. 150) as 'the fundamental form of self-knowledge', can thus give us (when successful) the relation between immediacy and first-person authority via the subject's use of transparency as a rational agent in a practical, reasons-based account of self-knowledge. Moran expresses his view in the following way:

> the primary thought gaining expression in the idea of 'first-person authority' may not be that the person himself must always 'know best' what he thinks about something, but rather that it is his business what he thinks about something, that it is up to him. In declaring his belief he does not express himself as an expert witness to a realm of psychological fact, so much as he expresses his rational

[32] Knowing something that you believe to be the case only tacitly is discussed in Chapter 4. Here we need to say just that such a belief is known only tacitly because it cannot be brought to the subject's full, explicit, conscious awareness. Abigail cannot admit *to herself* that she is ashamed of her obesity and that this is why she envies her friend.

> authority over that realm. (2001, pp. 123-124)

To conclude this discussion of how the transparency condition requires justification rather than explanation, I now set out another quotation from Moran and then relate it to the Caesar crossing the Rubicon example that now needs rounding off. Here is the quotation:

> The transition from attribution to avowal is thus an expression of the person's rational freedom, an assertion of authority. It is this assertion, this commitment, that makes it possible for her declaration to conform to the Transparency Condition, the announcement of her belief without reliance on the psychological evidence about herself. At the point, the proposition itself gains admittance as a basis for her further thought, both practical and theoretical. And this expresses the dominance in her self-reflection of justifying reasons over explanatory ones, for the premises of reasoning are not propositions about someone's beliefs but propositions about the objects of one's beliefs, the facts themselves. This expresses a relation to one's state of mind that is exclusively first-personal and not shared by the best telepaths of our philosophical imagination The agent belongs at the centre of the picture because if I am able to avow my belief, and thus speak my mind without the support of evidence about myself, it is because, within that context anyway I am taking what I believe to be up to me. (Moran 2001, p. 151)

This quotation makes it clear that in our earlier example of the subject who knows that Caesar crossed the Rubicon because he learned this fact from an excellent history teacher, the 'because' explains his knowledge in a third-person way but does not justify it. This is because his knowledge was not 'up to him'. He did not use his own reasoning power to decide that Caesar crossed the Rubicon. He can now successfully attribute to himself his belief *that* Caesar crossed the Rubicon, from the third-person point of view, and he can explain why he knows this (he learned it from his teacher). But he cannot *avow* either that or why Caesar crossed the Rubicon because he did not make up his own

mind about these by using his own reasoning power. His belief traded on his teacher's knowledge. It was not up to him. For this reason, his self-ascription does not conform to Moran's transparency condition. As Moran says in the above quotation, it is a statement of what is in his mind about Caesar; it is *not* a statement *about the facts,* made by using his own reasoning power, because he has not made up his own mind for himself what those facts are. It is not, therefore, an avowal at all.

Differences between Moran's account of transparency and empirical transparency accounts

As I have set it out above, Moran's use of transparency differs from empirical transparency accounts in at least two ways.

Firstly, Moran argues (p. 62) that even as Edgeley and Evans (separately) discuss it, they may be using the empirical fact of transparency to make a conceptual point about how we acquire self-knowledge *as agents* as well as the more general point about the relation between self-knowledge and the world. This conceptual point is that the idea of answering the question whether *p* is true (and thus whether one believes that *p*), where *p* is some external matter, cannot make sense to us *unless* we answer it by attending to the external facts that make it true. It is only when the subject adopts Moran's deliberative stance that she *can* answer the question whether she believes that *p* by considering exactly the same phenomena, and in exactly the same way, as she would in answering the question 'Is *p* true?' This is not the case when we are considering an empirical question about someone else, such as whether Jones believes that *p*. We can conclude that Jones believes that *p* by listening to Jones talking about *p*. *We* do not have to consider *p* at all in order to draw this conclusion. This provides further evidence of the asymmetry between the first-person and the third-person stances.

In discussing transparency as Moran uses it, therefore, we are talking only about situations where the subject avows her belief from the deliberative stance. She cannot do this unless she has deliberated upon the question 'Is *p* true?' (where deliberation is required) and has found her answer in doing so. When she is in the theoretical stance, transparency does not apply. This means that when we consider it in connection with the immediacy and the authority of self-knowledge, we cannot treat transparency as *only* an empirical fact (although it is, of course, an empirical fact), since we can make sense of our belief that *p* only by adopting the deliberative stance to determine empirically whether *p* is true (2001, pp. 62 to 63).

Secondly, Moran makes it clear that forming a belief for oneself via transparency does not necessarily amount to knowing that one has it, although normally it does amount to this. He is careful to explain that forming it for herself does not *in itself* produce the subject's on-going self-awareness of her belief: 'my account does not attempt to explain how the deliberative forming of one's beliefs produces self-awareness of the beliefs formed, because I don't think that's true' (2004, p. 468).

Not all beliefs, just formed, become standing beliefs, available to conscious awareness in relevant situations. Some get lost as soon as they are formed or later, and the subject may never know that she had ever formed them. So Moran is not explaining immediate knowledge of our intentional attitudes in general in his transparency account. He is saying only that *when,* while she is in the deliberative stance and focused solely on the belief's external object, the subject forms a belief whose sole reasons for her having it are transparent to their object and make sense to her, *and she sincerely avows* this belief, *then* we can say that her avowal of it justifies the authority of her immediate knowledge of it. The avowal has the authority of *her reasons* because she has formed the belief herself in conformity with the transparency of her belief to

the truth, as she sees it, of what she believes, and she knows what these reasons are because she has assessed the evidence for them and formed them for herself in a self-aware way.

Immediacy and transparency stand or fall together

One of Moran's major claims about the authority/estrangement relation is that immediacy and transparency go together. In his account, first-person authority and transparency are conceptually connected: 'immediacy and transparency, understood as describing the conditions for first-person authority, stand or fall together' (2004b, p. 457).

By 'immediacy' here, Moran means knowing the attitude that you have created *for yourself*. He is talking about more than just the conditions for the first-person authority of self-knowledge. He is also saying that when transparency and immediacy fail, we do not have something minor that we can just call anomaly. Instead, we have something that is as important to an account of self-knowledge as is first-person authority itself: we have, in his words, alienation or estrangement. He marks the importance of this claim by putting authority and estrangement on equal terms in the title of his book.

He also says, 'I do mean to show how the conditions that make for the assumption of first-person authority also provide the standing possibilities for certain characteristic forms of alienation' (2001, xxxiv): that is, he wants to show that the same conditions (immediacy and transparency) that produce first-person authority when they are met also produce its opposite, estrangement (alienation) when they are not met. A key to understanding self-knowledge, on this view, lies in understanding the relation between its authority and the lack of it in estrangement. Estrangement gives us cases where a subject has a belief that seems to be inherently unstable: that is, unstable not because the subject lacks key knowledge about its object but because even when he has all the

knowledge about this object that there is, his reason for having the belief remains opaque to him because an aspect of this knowledge is inaccessible to his conscious awareness. Estrangement is a phenomenon that merits serious attention; it deserves its own theoretical place in any account of self-knowledge. Indeed, Moran goes further. He argues that it is because of the link between authority and estrangement, provided by his two stances together with the transparency condition as he uses this, that we can have a practical account of self-knowledge *as achievement*.

This idea, that self-knowledge is an achievement, is tied to the further ideas that we can have attitudes that we are unaware of having because we are motivated to avoid knowing that we have them: that is, attitudes that we are estranged from. Barnaby denies that he is in love with Clarice when all their friends can see that he is. He is motivated to make this denial because he is afraid that if he admits his love for her, Clarice will reject him. But his fear of rejection and his love for Clarice are both opaque to him. He is estranged from them: he has no conscious, *immediate* knowledge of them. Barnaby may be self-deceived. If he insists that he likes being with Clarice only because they are both interested in heritage issues, he is self-deceived. You intend to eat a chocolate bar while knowing that this is against your own better judgment. You think you know why: 'It's because I can't resist chocolates', you say. But you do not know *why* you cannot resist chocolates. If eating them is against your better judgment, explicitly known at that time, you may be acting akratically. I argue in Chapter 6 that serious akratic action always involves a degree of estrangement such that the subject suffers a degree of psychological unfreedom.

Transparency is an achievement because it sometimes involves becoming aware of our estranged attitudes, a process that involves resolving inner conflict. Moran also describes it as an ideal, because our human capacity for

rationality is 'partial, fragile and imperfect' (2004b, p. 468). He comments at the end of his answers to objectors in 'Replies to Heal, Reginster, Wilson, and Lear' (2004b) that transparency as he conceives of it is an idealisation, distant, in one way, from how we really are (p. 472). On p. 468 of 'Replies' he adds that one's capacity to form an attitude on the basis of reasons is easily compromised. He makes the same point in *Authority and Estrangement* (2001, pp. 62–63), where he says that when one's belief that *p* is transparent to the truth of *p*, this is not a matter of first-person logic but is rather 'a kind of normative ideal: 'I see conformity to Transparency as an *achievement* of the person, the satisfaction of a normative requirement (2004b, p. 461, author's emphasis).

He also says:

> But all of this, the immediacy, the transparency and the authority, will only be in place to the extent that the person's reasons really do determine what his beliefs and other attitudes are. That's why I said that they stand or fall together … And I take it that this assumption, while indispensable to understanding someone as a believer at all, refers to a human capacity that is partial, fragile, and imperfect (2004, p. 457).

Thus, successful conformity to transparency in avowal is the norm, and this norm *is* the ideal. It is the normal, everyday use of avowal and it succeeds most of the time. Even so, there are times when avowal is hard to achieve.

Even though millions of avowals achieve transparency every day with relative ease, their 'ordinariness' needs to be seen against a background of a multitude of other cases where estrangement, rather than the brute imperfections of rational agency, prevents this. Looked at this way, avowal is an achievement when we set it against all the cases where we do not really know our own minds at all because, without consciously knowing it, we suffer from an estranged attitude towards the topic at hand. In this way of looking at it, ordinary, everyday avowal is an achievement whenever it succeeds in having first-person

authority. It is an achievement because being able to look squarely at whether *p* is true without any dissembling, conscious or unconscious, is an achievement of active, rational agency every time it occurs. This is not because it is so hard to reason correctly even when estrangement does not occur, but because i) it is so often hard for us consciously to set aside our prejudices and ii) we have so many prejudices! Did South Sydney score that goal? X, a South Sydney supporter, believes that they did. But Y, who supports the Sea Eagles, believes that there was a forward pass involved and no goal was scored. The video replay is inconclusive as to the forward pass. But to X and Y it is quite conclusive. X is confident that the pass was not forward and Y is equally confident that it was. 'Look', says Y. 'You can *see* that it was forward.' But X can see that it was not.

If this interpretation is correct, then when Moran declares that our human capacity for rationality is partial, fragile and imperfect he does not mean just that we make brute errors. He means that our rationality is easily compromised by our addiction, conscious and unconscious, to our many and various biases.

Many of us have standing prejudices about certain cultures, religions, races, political parties, sports clubs, schools or even the family down the road. But we may also have friends in these places. We may become estranged from our prejudice in some situations. Sometimes we put it aside, and make an avowal that has complete first-person authority. At other times the prejudice motivates us, consciously or unconsciously, to make avowals about our friends that have no first-person authority. Moran's transparency condition covers these cases. On his account, estrangement and first-person authority always oppose each other. One reason why first-person authority is so hard to achieve is that so often we are simply not aware, at the time, that we feel a certain way about the person we are discussing, even though at other times we can acknowledge this.

In this section, in setting out Moran's transparency account, I have discussed the relation he argues for between first-person authority and estrangement. In the next section I set out the relation he argues for between estrangement and the anomalies of self-deception and akrasia.

Section 4: Estrangement and the anomalies

Although a subject's first-person, present tense self-ascription of her intentional attitude normally has the authority of fully active, first-person agency, there are some cases where, because she is estranged (alienated) from her real reason for her intention, or (where no estranged attitude is involved) because of thoughtlessness or recklessness fuelled by emotion, the subject suffers a lack in her wholehearted knowledge of why she acts as she does. In this kind of situation she must therefore act with a degree of passivity concerning it. The lack of wholeheartedness is caused by the conflict that she must suffer when she does $\sim a$ while explicitly knowing, even as she acts, that she ought to do *a*. The passivity reduces her capacity actively to control her decision to do *a* on the basis of reasons. If no estrangement is present, she may not make sufficient effort, through thoughtlessness or recklessness or because the emotion driving her and producing conflict leaves no room for sufficiently considering what she is doing. The result is that her self-ascription, although agential, is not fully self-determined. If an estranged attitude is involved, she necessarily cannot act freely. If no estranged attitude is involved, she still does not fully know her own mind: her self-knowledge is flawed. In both cases, her self-ascription has only third-person authority; she has not achieved normal, first-person authority over it.

Some writers acknowledge attitudes that are inaccessible to conscious awareness. For example, self-deception, according to Audi 'is primarily a state in which a kind of psychological dissociation gives rise to a disparity between

what the self-deceiver knows, albeit unconsciously, and what he avows or is disposed to avow' (1988 p. 117). Here we have an acknowledgement of the concept of unconscious self-knowledge (knowledge from which the subject is dissociated/ estranged/alienated) and of a disparity between this kind of self-knowledge and self-knowledge of which the subject is consciously and explicitly aware at the time she acts. Unconscious (estranged but tacitly known) mental states *interfere with* rational agency, often causing the subject to produce a false rationalisation masquerading as her real reason for her belief (as in self-deception). Such estranged attitudes are *potentially* available for later recall to conscious awareness.

Some accounts acknowledge the presence of conflict in the subject where an anomaly is present. Stroud and Tappolet, for example, tell us that philosophical discussions of akrasia have focused

> not so much on weakness of will as a character trait as on the sort of action that manifests it: roughly, intentional action contrary to one's better judgment: that is, contrary to the judgment that another course of action would be better. (2007, p. 2)

Stroud and Tappolet do not refer to unconscious attitudes, but their view is consistent with positing such attitudes to the extent that it presents, as anomalous, cases of akrasia where the subject is conflicted because of the involvement of an inaccessible (estranged) attitude. In akrasia, the conflict is between what the subject believes she ought to do, all things considered, and what she does. Both estrangement and conflict, on Moran's account, are present in self-deception, between self-knowledge that the subject is consciously aware of and self-'knowledge' that is not her real reason (her real reason has been repressed). If we accept that unconscious self-knowledge can conflict with conscious self-knowledge, so motivating anomalous behaviour, it is reasonable to suggest that sometimes the reason for the conflict is that there

is an estranged attitude motivating the subject. At other times, because of thoughtlessness or recklessness, often fuelled by emotion, the subject suffers a lack in her wholehearted knowledge of why she has this attitude. But in both cases, we need to postulate relevant unconscious self-knowledge in the subject, because it is unconscious self-knowledge that blocks her off from knowing what she *really* believes, feels or wants to do.

As we saw in the last section, Moran argues that estrangement prevents transparency. Transparency in turn, he also argues, 'is more of an achievement than something with a logical guarantee' (2001, p. 67). He relates the difficulties of achieving transparency directly to the anomalies, by saying that it is the cases where estrangement applies that make the achievement of transparency impossible, and that this is particularly the case in akrasia and self-deception. In self-deception and in serious cases of akrasia, he argues, my belief that *p* is not 'up to me', because although I may know *that* I believe that *p*, I do not know *why* I believe that *p*. This knowledge is inaccessible to my conscious awareness and is thus opaque to me. It is tacit as a standing attitude but it differs from ordinary tacit knowledge in that it cannot be brought to consciousness when required. It might be, for example, 'some anger or fear persisting independently of my sense of any reasons supporting it' (2001, p. 67). I feel the fear but I cannot explain to myself why I have it; there seems to be nothing to fear. This applies especially to phobias, such as fear of spiders.[33]

These sorts of situations, Moran comments, help us to see that the theoretical certainty that deflationary accounts use is not a satisfactory model for the achievement of self-knowledge. It is the practical ability to endorse one's belief

[33] I do not discuss phobias in this thesis because they involve non-conventional symbolisation, and there is no space to explore this. In general I think phobic objects usually symbolise attitudes from which the subject is estranged.

as true that is more important. In self-deception and akrasia, the subject cannot endorse her belief as true or her intention as choice-worthy. The serial gambler cannot endorse her belief that she will not gamble again, a claim that for her can be no more than a theoretical prediction. The mother of a daughter who is doing badly at school cannot admit to herself that her daughter actually has learning difficulties because this would be too shameful. So she rationalises (invents the excuse) that it is the teacher who is to blame for her daughter's poor results. However, she cannot endorse her rationalisation as true by producing evidence for it. She must insist on believing something that is contrary to the evidence.

Moran's transparency condition itself is evidence that neither self-deceptive nor akratic acts have first-person authority. The mother whose daughter has learning difficulties cannot self-ascribe either 'My daughter does not have learning difficulties' or 'It is my daughter's teacher who is the problem' with first-person authority because both claims are fuelled by her estranged attitude of shame. This attitude of shame is something about her, not something about her daughter. Her self-ascription is about her daughter, but her reason for self-ascribing it is about her.

Conclusion

In this chapter I have described and related to each other the main concepts in Moran's account of self-knowledge. I have placed estrangement within the structure of concepts that are basic to this account, and I have shown how the asymmetry between the first- and the third-person perspectives on self-knowledge underpins the commonsense realism of his account as a whole. I have set out the Kantian basis of Moran's rational agency account and have then described his transparency account, showing how this can give us a way of separating out those cases that have full first-person authority from those

that have a lesser degree of this, and why first-person self-knowledge is substantive. Finally, I have shown how Moran links estrangement to the anomalies of self-deception and akrasia. On his account, as I interpret it, cases of self-deception and of serious akrasia always involve estrangement, and estrangement implies a degree of psychological unfreedom in the subject.

But before discussing self-deception and akrasia, I defend Moran's account against objectors. In Chapters 2 and 3, I defend it against deflationary objections to his rational agency and transparency accounts. In Chapter 4, I show how Boyle's reflectivism supports Moran's position. I turn to self-deception and akrasia in Chapters 5 and 6.

Chapter 2: Rational agency

Introduction

In Chapter 1, I set out the basics of Moran's account of attitudinal self-knowledge, concentrating on the notions of asymmetry, rational agency, transparency, estrangement and the anomalies. I move now to a three-chapter discussion (Chapters 2, 3 and 4) of objections to Moran's account. This first chapter of the three is short. Its aim is to defend his account against two criticisms of his concept of rational agency.

Moran's rational agency account has been subjected to a range of criticisms. Two standard criticisms are that it is too rationalistic and that his concept of first-person avowal is reducible to that of third-person attribution. Owens (2003) and Carman (2003), for example, have both criticised it for being too rationalist and Owens has also attempted to show that avowals are symmetric with third-person attribution in cases of an intention to *a*.

In Section 1, I consider Owens' objection regarding intentions. I conclude that Owens fails to make his case because he fails to take account of Moran's first-person /third-person distinction, underpinned as it is by the Kantian active/passive distinction with respect to agency.

In Section 2, I turn to Carman's objection that Moran's account is too rationalistic. Carman claims that to have first-person authority on Moran's account, intentional attitudes must be justifiable by 'cognitively articulated' reflection or deliberation: that is, the subject must be able to give some reason why she has avowed her belief (or other attitude) that *p*. Carman holds this view because he believes that rational reflection is a requirement of being answerable to reasons. I argue that answerability to reasons, as Moran

explicates it, requires cognitive deliberation in many but not all cases.

Our basic question in this chapter, therefore, is whether Moran's rational agency account is consistent with a subject's first-person, substantively epistemic and justifiable knowledge of her intentional attitudes without being excessively rationalistic and while remaining asymmetric[1] with the third-person perspective (recall from the introduction that I follow Moran in holding that two perspectives are symmetric if self-knowledge attributed from both are equally authoritative; in this chapter, I will understand speech acts as symmetric if they satisfy this same condition).

I argue that the answer to this question is yes.

Section 1: Owens' objection: 'avowed' intentions are reducible to attribution

Owens' objection attempts to show that where intentions are concerned, avowal is symmetric with attribution. This objection uses as its example the case of the Catholic woman made pregnant by rape. Owens argues that although this subject cannot avow her intention to have an abortion, she can affirm it. Her affirmation gives her knowledge of and control over her intention and yet does not transcend the empirical statement of what her intention *is*: thus, where intentions are concerned, affirmation plays the same role as avowals do.

If Owens is correct, his argument puts Moran's notion of avowal under pressure because, it would seem, 'what manifests a distinctive knowledge of (and

[1] 'Being asymmetric with' might not necessarily involve 'being reducible to' the third-person, attributive stance in the sense of having the same authority as that stance normally has. However, I do argue that it is this kind reduction that Owens is aiming at and that his intention along those lines fails.

control over) intention is not our ability to avow these intentions but rather our ability to affirm them' (2003, p. 795). Owens is claiming that in general there is a difference between how we can know our beliefs and how we can know our intentions. This looks reasonable. Moran himself says that although in his 'Precis of *Authority and Estrangement*' (2004a) he talks mainly about beliefs, other intentional attitudes such as intentions, desires and affective attitudes may need different treatments.[2] The difference Owens is interested in is that for beliefs to have avowable, first-person authority, the subject must endorse them as being true, an act that transcends the empirical fact that the subject *has* the belief to consider the normative question of whether she endorses it as true. Intentions, on the other hand, require, for the subject's knowledge and control over them, only an empirical affirmation of her intention. Thus avowal, a first-person concept on Moran's account, is symmetric with affirmation, a third-person concept, where intentions are concerned. I argue now that the Catholic woman does not have control over her affirmation.

The Catholic woman

Speaking of the case of te Catholic woman intending to terminate a pregnancy produced by rape, Owens says,

> Consider a Catholic woman … who has become pregnant after rape. She judges that she ought not to have an abortion but akratically decides to have an abortion nevertheless and makes plans to attend the abortion clinic next week. Can she avow this intention? She can do more than report this intention in the way she might report a third party's intention but she cannot go so far as to endorse this intention. What she can do is affirm that she is set on having an abortion, that she has resolved to have an abortion, that she is committed to having an abortion. Moran's notion of an avowal seems to include the idea of endorsement (2001, p.

[2] See Moran 2001, p. 123.

> 67), so perhaps he would maintain that the woman cannot avow her intention. But if that is how the notion of an avowal is to be understood, it looks as if what manifests a distinctive knowledge of (and control over) intention is not our ability to avow these intentions but rather our ability to *affirm* them (p. 795).

This Catholic woman certainly seems to be doing more than affirming her intention. In Owens' own words, she is self-ascribing it *with commitment and resolution*. She is '*set on*' it (Owens' words). This suggests that she is at least trying to endorse it, and thus is speaking from Moran's deliberative stance, rather than merely reporting her state of mind from his attributive stance, as if she were reporting someone else's intention.

But having claimed that the Catholic woman is committed to and set on having the abortion, Owens also claims that she cannot endorse her intention to have an abortion because, being a good Catholic, she knows she should not have one. Thus she cannot avow her intention. She can affirm only that she intends to have an abortion, and her affirmation does not amount to avowal because she cannot justify it, even to herself. But Owens claims that it is more than merely a third-person, attributive statement that she *has* that intention; it is an affirmation of it that shows knowledge and control over her intention. This, Owens suggests, shows that there is a difference between beliefs and intentions where their practical authority is concerned. Where our beliefs are concerned, we need practical endorsement as well as empirical knowledge of the belief. But where intentions are concerned, affirmation, he argues, already gives us knowledge and control over our intention.

However, because the Catholic woman cannot fully commit to and resolve to carry out her intention, she in fact does not have complete knowledge and control over it, regardless of what she says. Even as she has the abortion, her feeling of alienation from what she is doing indicates that she is acting with a degree of passivity concerning it. If we compare this with the situation of the

drug addict who voluntarily injects his arm with the drug even while hating himself for doing so, it might become clearer that although conflicted akratic subjects sometimes *seem* to have full knowledge and control over their intention, they do not really have such control and thus cannot successfully avow it. They can, of course, affirm it. But their affirmation cannot be wholehearted or settled; their intention cannot be fully self-determined.[3] Owens' conclusion is that a subject who deliberately fails to do what he thinks he ought to do is not less well placed to know what he is doing and why than is someone who does what he thinks he ought to do (2003, p. 797). I conclude that this is incorrect.

Conclusion

In this section I have discussed Owens (2003) objection against Moran's (2001) notion of rational agency that first-person agency is reducible to the third-person, empirical perspective in the case of intentions because the subject (the Catholic woman made pregnant by rape) knows and controls what she intends to do just by affirming rather than by avowing this intention. I reply that affirmation does not imply *first-person* authority, because, again, the subject who can only affirm rather than avow her intention is conflicted.

I conclude that Owens' attempt to reduce first-person to third-person self-knowledge in the case of intentions fails. In order for knowledge and control to be present, a self-ascription from the first-person perspective requires the subject's wholehearted endorsement of and commitment to her intention as well as to her belief.

---

[3] Where the subject cannot endorse either of two alternatives, her self-ascription does not have active agency because at that time she cannot fully identify with what she intends to do, even as she does it. She cannot feel that her intention is fully *hers*; it does not have her full and active commitment to it.

Section 2: Carman's objection: excessive rationalism

In this section I discuss Carman's objection. Carman, like Owens, objects that Moran's account is too rationalistic, arguing that he relies on an overly intellectualised conception of intentional attitudes. According to this conception of them, Carman claims, 'all such attitudes, precisely in order to be first-personal in the proper sense, must be responsive or answerable to reason' (2003, p. 400).

To be answerable to reason, such attitudes require cognitively articulated reflection or deliberation, which, Carman argues, is often unnecessary. We cannot, he says, 'reduce the entire first-person practical standpoint to the sphere of rational reflection … I don't know why we should conceive of (Moran's) first-person perspective solely, or even primarily, in terms of rational reflection' (2003, 403; 404). I argue here that Carman's position represents a misunderstanding of Moran's (2007, p. 54) distinction between active and passive rational agency.

Examples that Carman gives to support his position are mostly of what he calls 'mundane pre-reflective forms of understanding' (2003, p. 399). These, he claims, include i) beliefs not arrived at by explicit deliberation and ii) non-cognitive attitudes such as desires and emotions, caricatures and stereotypes, and sub-rational moral self-interpretation. I begin with i) beliefs not arrived at by explicit deliberation.

Beliefs not arrived at by explicit deliberation

Carman argues that Moran reduces 'first-person practical attitudes at large to those that are specifically reflective or deliberative', and suggests that 'all such attitudes, precisely in order to be first-personal in the proper sense, must be responsive or answerable to reason' (2003, p. 400).

Moran's reliance on the answerability of avowable attitudes to reason is certainly central to his account. One reason is that in his view, our intentional attitudes are answerable to how the world is. This is Moran's commonsense realism about mental states—that at bottom, they are determined by the world. Wanting to be a barrister is a good reason for wanting to study law because the world is such that there are law courses that can qualify you to become a barrister. This is how we know that if this is your real reason for wanting to study law then your desire to do so is rational—but if your real reason is to please your father but you sincerely believe that it is to become a barrister, then you are estranged from your real reason and you may be self-deceived.

It is also true, of course, that many of our beliefs are not arrived at by explicit deliberation. Moran's claim is that we expect agents to take responsibility for their beliefs with respect to how they behave when action concerning them is required. In this sense reasons are involved in perceptual beliefs, remembered beliefs, and beliefs adopted on someone else's say-so, even though these may never require deliberation. Being engaged with reasons is about adopting a practical stance towards one's belief, not about trying to conform to a normative demand that all agents ought to be consciously aware at all times of the need for occurrent episodes of their attitudes to be rational. However, this engagement with reasons does imply, for Moran, a different normative demand. All our beliefs form a network, and as agents we ought to do something when some part of this network is challenged or abandoned. As agents we ought to justify or correct one or more of our assumptions. Without some assumption of agency there would be no point to this demand for justification or correction, for there would be nothing we could be expected to do in *response* to this demand (Moran 2003, p. 404).

Reason, Moran claims, is engaged in a wide variety of pursuits, including

practical tasks such as finding one's way in a complex social situation, following the music in dancing, coping with some problem and using attention and critical discrimination in pursuit of a goal: 'as long as there is room for the idea of correction and getting back on track, there is room for the idea that the person has reasons for going this way rather than that way' (Moran 2007, p. 54). The subject must *have* a reason for doing what she does, that she can articulate when appropriate. The dancer, for example, is trying to follow the music, to get it right, to keep in touch with it. We know she is engaging her reason in doing this because she can get the dance steps wrong and then correct them.

Moran also points out that the idea of reason must include the ideas that the avowed belief, intention or affective state[4] is in some way objectively choice-worthy, that reasons come in many different varieties and grades of value 'and that these are not all commensurable with each other' (2007, p. 64). Being not all commensurable with each other suggests that although, if reason is being engaged in these attitudes, there must be some connection between the attitude itself and 'some assessment of the characteristics of its object' that make it true or choice-worthy, Moran wants to interpret this connection very broadly to include cases where the attitude conflicts with other values of the subject (p. 65). He says, for example, that 'I just felt like it' can count as justifying the subject's goal as 'in some sense worth pursuing'. Carman's claim that reason may sometimes play no part in justifying avowable self-knowledge can thus be addressed at a general level by saying that demanding a reason is sometimes inappropriate, such as when we look out the window and exclaim, 'It's raining!' It may be inappropriate in other situations, too. For example, Nick

[4] Not all affective states have an object—I can just feel happy. Only affective states that do have an object can be incorporated into Moran's account of self-knowledge.

Kyrgios may serve towards the centre of the court on a split-second decision to do so. He may not be able to justify this move later to his coach. But we can assume that he did have a reason at the time and that he could have enunciated this reason were we able to freeze time at that point while we asked him. Such cases do not necessarily imply any lack of first-person authority over his intentional action. However, when a justifiable reason is appropriate, the subject must be able to give one that is objectively minimally rational and conforms to transparency when the context of her avowal is fully unpacked. As I argued in Section 1, people do not avow that *p* from the deliberative stance for no reason.

Cognitively unarticulated beliefs

Carman also challenges Moran's position in connection with cognitively unarticulated beliefs: 'if my "guiding reason" is something cognitively unarticulated, such as an emotion or a desire, then I can indeed know what I'm doing, and even why I'm doing it, though reason plays no positive role in either generating or justifying my action' (Carman 2003, p. 405). Reason, he is saying here, may in some cases play no part in justifying self-knowledge gained in Moran's deliberative stance. Even some explicitly deliberative attitudes might be non-cognitive:

According to Moran, then, genuinely constitutive self-interpretation would require that all the attitudes at play in the process be cognitive attitudes, for 'a new description of my emotion or belief is powerless to alter it unless I believe the description' (Carman 2003, p.55). Merely adopting new descriptions, interpretations, formulations or vocabularies, Moran thinks, implies a kind of arbitrariness, if not outright voluntarism (p. 402).

Moran argues, Carman says, that all constitutive self-interpretation must involve belief; anything else would be voluntarism. But there is no reason,

Carman argues, to limit to cognitive attitudes the process by which we constitute our mental states. He points to caricatures and stereotypes: 'They do not advance explicitly articulated "belief-like" claims, but instead simply present their objects as, say, sinister or grotesque or ridiculous. They do not assert anything; they just make things appear in a certain (favourable or unfavourable) light' (2003, p. 402).

However, although caricatures and stereotypes may not cognitively articulate anything overtly, they do communicate certain ideas about possible evaluative interpretations of what is being caricatured or stereotyped. Carman acknowledges this himself by commenting that they 'make things appear in a certain … light' (2003, p. 402), thus raising questions about attitudes for the viewer to consider and form beliefs and affective attitudes about when these are required or appropriate. Here we have a difference that Moran points to, between being a product of deliberation and merely being answerable to deliberative considerations (Moran, 2001, p. 63). This example is consistent with such attitudes being answerable to reasons. When the need arises, a cartoon depicting a political figure in a certain light with respect to a recent situation or event is answerable to our judgment as to (say) its appropriateness.

Having first-person authority in a situation where you are answerable to reasons must be differentiated from having first-person authority when you have an estranged attitude. The racist who believes black people are stupid because he hates black people does not have first-person authority over his self-ascription 'I believe black people are stupid' because his hatred of black people is estranged from his conscious awareness of it. His 'avowal' (if that is what it is) has failed. From the third-person point of view, it has only the epistemic authority of the third-person attribution. He cannot justify it if asked and it does not conform to transparency because he has a reason for making it that is about himself (that he hates black people) rather than about black

people.

Moran also implies, Carman points out (2003, p. 405), that in the absence of reasoning about what I am doing I cannot know why I am doing it. He quotes Moran:

> The description under which an action is intentional gives the agent's primary reason in so acting, and the agent knows this description in knowing his primary reason. This description is known by him because it is the description under which he conceives of it in his practical reasoning. (2001, p. 126)

Carman takes this to imply that in the absence of reasoning, when my guiding reason is cognitively unarticulated, such as when it is an emotion or a desire, I cannot know the *reason* (his emphasis) why I'm doing what I'm doing (2003, p. 405). He points out that this does not imply that I do not know what I'm doing or why I'm doing it or that I cannot justify my action even though reason plays no positive role in either generating or justifying it (p. 405).

But if you cannot say, when asked, how you know it is raining outside, if you cannot say, 'because I looked out the window' or 'I heard it drumming on the roof', or even 'It always rains here at four o'clock', then why would we suppose that you have any first-person authority about your belief that it is raining? It seems rather that you might have decided to believe it is raining quite randomly, in a voluntaristic way. But belief is not voluntarist. We 'decide' to believe that *p* by judging that *p* is true. Of course, there are cases intermediate between voluntarism and being able to state one's reasons. I tell you that it is raining. You ask me how I know. I confess I don't know. But there is a reason: I've forgotten that I saw mud on the carpet.

Non-cognitive attitudes such as desires and emotions, Carman also argues, do not require rational foundations. Let us consider these separately.

Desires

Carman describes an example of what he calls sub-rational moral self-interpretation in Plato's *Republic*, Book 1V, where Leontius lustfully desires 'to look at a bunch of corpses, feeling ashamed of the appetite, but making no rational judgment that it would be wrong to look' (402). Here, he says, the diminished role of rational judgment is crucial to Plato's argument, which purports to demonstrate inner psychic conflict between moral emotion and appetite rather than between reason and appetite. But when we unpack this we find that there is nothing sub-rational about this example.

Leontius desires to do something concerning some external objects—to look at them. His desire is lustful—let us say it excites him that these people are dead. Perhaps they were his enemies in a recent battle. But he avoids judging that it would be wrong to look.

The desire to enjoy this looking is his reason for wanting to look at the corpses. It might or might not be a morally acceptable reason—he does not consider this. In fact he avoids considering it, so that he does not have to worry that feeling pleasure at their death might not be morally good.

This implies that he feels ashamed for a reason that he is currently estranged from. It is a reason that he *does have*; otherwise he would not feel ashamed. So he can *report* that he feels ashamed of his lust but he cannot avow this. This is a not uncommon situation. He wants to look at the corpses for a reason that he is aware of (it will give him pleasure) but he avoids considering any rational judgment as to the rightness or wrongness of enjoying this looking because right now he does not want to acknowledge his belief that enjoying this pleasure might be morally wrong. He is temporarily estranged from this latter belief; I agree with Plato that he suffers moral conflict. This does not imply any lack of his answerability to rational judgment to some minimal extent that can be called on when appropriate. But it does imply that his intention to look has only third-person authority.

Carman's argument that not everything must be responsive to reasons thus cannot be substantiated; Leontius suffers moral conflict *for a reason*. Carman seems to think Moran is saying that we must constantly be making conscious rational judgments about what we should do, whereas Moran acknowledges that we often avoid noticing the moral or rational status of what we are doing. This avoidance introduces a passive element into our thinking. We necessarily cannot fully and wholeheartedly endorse our intention to do something while it continues to operate.

Emotions

Much the same reply can be made to Carman's objection concerning emotions. He argues that emotions, also, do not always need rational foundations and that some emotions, such as shame and forgiveness, have an essentially non-rational dimension: 'We often have reasons to forgive or not to forgive, but purely rational considerations can never capture the unique intelligibility of forgiveness' (2003, p. 403). But Moran does not claim, and nor does his account require, that emotions can be *reduced* to reasons, only that they can *function as* someone's reason for what he does. For example, you might forgive another because you love him. Many would say that love is a good reason for forgiving someone. But the point is that you *have* a reason, both for forgiveness and for shame.

Carman also points out that we can express an emotion without necessarily making a judgment about its object. But without making a judgment about our emotion we are not subjecting it to reason:

> 'I can feel ashamed without judging that I have done anything shameful, just as I can feel joy … without judging anything to be particularly joyous. Boredom is not, nor does it entail, a belief that something is boring (p. 402).

But this misrepresents Moran's position by assuming that it applies to

phenomenal as well as to attitudinal avowals. If your state of shame or boredom or joy has no object then we might call it a mood, and you cannot avow from the deliberative stance that you are feeling in a certain mood for no particular reason, because an intentional attitude must have an object. To say 'I feel joyful' is to take an empirical or theoretical stance towards your psychological state, a stance that is not answerable to normative assessment. It conveys self-knowledge, but is not answerable to Moran's transparency condition because it is not via the exercising of this subject's rational agency that he has discovered that he feels joyful—he feels joyful immediately, for no apparent reason. Wright (1988) explains that there are two broad classes of avowal: phenomenal avowals[5] and attitudinal avowals. Phenomenal avowals are groundless and strongly authoritative, but they have no object. Attitudinal avowals, on the other hand, are transparent to their objects (e.g., 'my foot hurts'). Moran restricts his account of self-knowledge to attitudinal avowals and in this thesis I follow him in doing so.

Moran's rationality requirement

Moran's answer to Carman's claim that we do not need a rationality requirement (2001, p. 63) is thus that there is logical room *in all cases of settled belief* for the question 'How do you know?' His rationality requirement thus covers all cases where a belief (even just the belief that it is raining) has been formed and is avowed. When the answer justifies the avowal ('I've just looked outside') the avowal has prima facie first-person authority.

Carman protests that this explanation is too weak, amounting only to saying

[5] Since Moran does not discuss phenomenal avowals he makes no claim as to whether they can be avowed or only self-ascribed. But 'avowal' as he uses it in connection with attitudes is a normative term, and a phenomenal self-ascription is not normative. It is in a different category.

that one can raise the question of rationality without threat of inconsistency (2003, p. 404). Moran's own claim is that it is the traditional way of relating rationality to self-knowledge that is too weak. As Child points out (2009, pp. 850–855), 'if all that rationality requires is that a subject knows *what* (my emphasis) she believes, the requirement could be met by someone who self-ascribed beliefs in the same way that he ascribed beliefs to others: on the basis of what he said and did'. Child goes on to say,

> But, Moran argues, such a person would be incapable of regulating his beliefs in the way that is distinctive of rational subjects: by reasoning about what to believe in a way that directly settles what he does believe. (p. 852)

The claim that knowledge of our intentional attitudes has authority seems not to require the rationality condition that Moran requires of it only if we think of such knowledge as something we just have, immediately, that is ours in the same way as our knowledge of our pain or joy is ours. But we have not formed our pain or joy by the exercising of our rational agency. We suffer or enjoy our pain or joy. Where our knowledge of the intentional attitudes is concerned, our immediate knowledge has first-person authority because we know the beliefs we have formed for ourselves in a special way (immediately), because we are in a special position to know these —they are our own. We know them differently not only from how we know the beliefs we have formed about others but also from how we know our *own* pain or joy, which we have not formed for ourselves by exercising our rational agency but nevertheless suffer or enjoy. Asymmetry between these different ways is intrinsic to knowing our own minds.

Moran's distinction between an active knowledge of ourselves as spontaneous beings and a passive aspect of this knowledge grounded in our power of

sensible receptivity[6] (2001, p. 114) is consistent with the claim that for first-person authority to succeed, the subject's avowal must conform to his transparency condition as outlined in Chapter 1 that is soon to be discussed in Chapter 3. When one's motive is about oneself rather than about the object of the avowal, transparency fails, and the self-ascription, from the third-person perspective of the listener, has only third-person authority, even when the subject is fully aware of it, if she has learned about it on the testimony of another rather than discovering it immediately, *for herself*. We saw this in the example of the Catholic woman who has to decide whether to have or to abort her rapist's baby. Discovering your reason about *X for yourself* means making up your own mind about *X*, so giving your avowal the first-person authority of active agency. Knowing *X* only empirically has a degree of passive agency because your active agency has been restricted or compromised by the fact that you do not know *X for yourself*. You have not entirely, wholeheartedly and in a settled way, made up your own mind about it. You cannot endorse its truth or choice-worthiness and commit to it in practice.

One point that Carman makes against Moran's 'excessive rationalism' actually illustrates the importance of the latter's stand on rationality. First mentioning that Moran (2001, p. 54) concedes that phobias do not seem to fall under rational criticism (they may remain even after the subject recognises that as fears they are baseless), Carman then points out that we might just 'decide to accept a person's idiosyncratic fears as brute responses, after all, and leave it at that' (2003, p. 401); making the person 'feel the pressure of reason … seems arbitrary at best, coercive at worst' (p. 401). In deciding 'how the interests of reason ought to figure in our dealings with this (p. 401) Carman says, 'What is

[6] Both Moran (2012 and 2001) and Boyle (2011, p. 224) link this distinction of Kant's with the transparency of attitudinal self-knowledge to its truth or choice-worthiness.

at stake is the applicability of the categories in question, not the mere desirability or undesirability of forcing the issue' (p. 402).

In fact we often do just accept that someone has a phobia and leave it at that. But this is not to accept that the phobia itself is rational. Being terrified of all spiders, from harmless to fatal, is irrational. And although forcing the issue would serve no good purpose for one who has a phobia, helping a phobic person to discover why he has it might. Helping him to discover why he has it, though, is to help him discover the real reason for his phobic behaviours. Discovering one's real reason is in the service of rationality; rationality is a tool that we need for self-understanding and thus for psychological health and self-growth.

Moran's account, by tying first-person authority to rational freedom and agency, provides logical space for fruitful explorations of opaque reasons and alienated, estranged attitudes and of self-deceptive, akratic and phobic reports that masquerade as avowals, thus providing a route towards explaining self-knowledge as an achievement worth striving for because it increases our well-being, rather than as something we have for free.

Conclusion

If Moran's account is viable, the concept of rational agency as he uses it does not produce an excessively rationalistic account of self-knowledge. On the contrary, self-knowledge can have first-person authority as long as the subject's fully active agency is being exercised rationally to a minimal extent, as long as this exercising of it is objective, and as long as the avowal conforms to transparency.

Moran's account opposes first-person authority to estrangement. Estrangement prevents the subject from acting as a rational agent by

obstructing her exercising of this capacity in her avowal of her intentional attitude in relevant situations. Unlike the dancer who can correct her wrong dance steps, the subject who has an estranged attitude either cannot consciously access it or cannot feel wholehearted about it. First-person authority comes from the fact that, normally, subjects *have* a reason for what they believe or intend to do; they know what this reason is because they have formed their attitude *for* that reason[7]; they *believe* their reasons are sound and we *expect* them to take responsibility for the actions that their reasons motivate.

The basic question in this chapter has been whether the rational agency dimension of Moran's account is consistent with the subject's substantively epistemic and justifiable knowledge of her intentional attitudes without being excessively rationalistic and without reducing first-person authority to third-person authority. I have argued that this dimension of his account meets both of these conditions. Estrangement, I will argue in Chapter 5, is the product of a form of repression (such as 'Freudian') that motivates the anomalies of self-deception and of serious akrasia. But first, we find in Chapter 3, as we discuss Moran's transparency account, that we must consider estrangement as the opposite both of immediacy and of first-person authority.

[7] This raises the question of how we know what we intend to do, and is discussed in Chapter 4.

Chapter 3: Transparency

Introduction

In this chapter I discuss objections to the transparency dimension of Moran's account of self-knowledge. The first objection, to be discussed in Section 1, is that transparency does not apply to pre-existing beliefs because to recall pre-existing beliefs we need to look inwards, into our minds, rather than outwards to the world. The objections of this kind that I discuss have been made by Gertler (2011) and Reed (2010). I argue that any empirical approach, such as the form of contemporary Cartesianism that Gertler espouses, can give us only a third-person theory of self-knowledge. If the first-person position is asymmetric with the third, such accounts fail to explain self-knowledge. More importantly, much of the opposition to Moran's view on pre-existing beliefs assumes that a self-ascription of such a remembered belief is made from his deliberative stance, whereas on Moran's account it is merely attributive. On his account, attribution does not use transparency.

Having considered Gertler's arguments, I oppose to them those of Peacocke (1998), Falvey (2000) and Moran (2012), who argue that pre-existing beliefs can be transparent. I then compare Reed's position with Peacocke's and Falvey's, to show that Reed's, Peacocke's and Falvey's amount to the same thing, although Reed starts from the quite different and incorrect premise that Moran's account is non-epistemic.

The second objection, discussed in Section 2, is that transparency cannot explain immediacy. This objection claims that self-knowledge is not immediate because it involves the use of inference. Even when, by using transparency, we judge that *p* is true and know that we have so judged, it does not follow from this, this claim says, that our knowledge that we believe that *p* is immediate. It

does not explain how we get from knowing that we judge that *p* to knowing that we believe that *p* without evidence or inference—in this case, without inferring from judging to believing. Shah and Velleman (2005), Cassam (2010) and Fernandez (2003) have all argued that the step from knowing that we judge that *p* to knowing that we believe that *p,* because it involves inference, is evidence-based rather than immediate.

These objections use the idea that we can try to find out what we believe about something without considering that thing itself (Moran, 2012, p. 223). The rejection of immediacy thus goes hand in hand with the claim that the subject not only *need* not use reasons in forming her belief, but also that she *must* not use reasons in doing so. This argument has been presented by Shah and Velleman (2005). I first discuss Moran's (2012) response to it and then give my own response to the deflationary accounts of Cassam and Fernandez. I argue, as Moran does, that these writers assume a concept of agency as being a process of production, that of acting upon oneself to produce a belief without the need for using reasons, rather than the quite different concept of agency that Moran himself uses, to which the subject's reasons for her belief are central.[1]

The objections discussed in Sections 1 and 2 are related in two ways. Firstly, both kinds of objection imply that the subject is using a passive rather than an active form of agency; she is not acting as an agent who is making up her own mind, for better or worse, about what she believes. Secondly, their common aim is to demonstrate that the first-person (reasons-based) position on self-knowledge is symmetric with the third-person, empirical approach. But quite

[1] Both Moran (2012) and Boyle (2011) criticise the 'production process' idea. Boyle argues, for example (2011, abstract), that beliefs are acts of reason whose capacity for self-determination lies in their very nature, not in any facts about how they originate.

apart from any other reasons, the third-person, empirical approach cannot provide us with the sort of conceptual structure that opposes first-person authority to estrangement in a way that can begin to solve the problem of the anomalies.

Section 1: Transparency and pre-existing beliefs

The objection from pre-existing beliefs is that such beliefs are not transparent because we look inwards into our minds, not outwards at the world, to remember them. Suppose that we want to remember what we had for breakfast this morning. Let us agree that we first adopt the attributive perspective in order to remember this. We do 'look into our minds'. On Moran's account we can report our belief that *p* from a third-person, attributive perspective, once it is already a settled belief of ours, without transgressing the transparency condition. This is because on his account, transparency does not apply to already known, settled beliefs. *This*, in turn, is because on his account the first-person avowable stance is asymmetric with the third-person, attributive stance. Empirical transparency accounts cannot allow this.

Common sense tells us that when we adopt the attributive perspective in order to remember what we had for breakfast this morning, we are probably going to remember something that we still believe. Suppose that we still believe now, having remembered it, that we had eggs for breakfast this morning. We can remember eating them. Moreover, we can still give reasons for having eggs. We like eggs, or eggs were on the menu. So when we say, having remembered it, 'I believe I had eggs for breakfast', we are saying something that we might be endorsing as true even as we say it; i.e., we might be avowing our belief. If we are avowing it, then we *are* using transparency, because, unless we are misremembering it, our belief, 'I believe that I had eggs for breakfast', is transparent to the truth of 'I had eggs for breakfast'. The third-person,

attributive perspective in this case *assumes* that we could also avow that same self-ascription if we wanted to.[2] But if we are still unsure about whether we really did have eggs for breakfast that morning, we can switch to the first-person, avowable perspective in order to check whether our memory is correct. In doing this checking, we are using transparency because we are now consulting the world. For example, we might ask our brother, who had breakfast with us: 'What did I have for breakfast?' This, it must be pointed out, is a question about *whether* I had eggs, not about why I had eggs. It is not relevant to the question 'Why did Caesar cross the Rubicon?' discussed in Chapter 2.

A better answer has been developed by both Falvey and Peacocke. I begin my discussion of these writers' arguments with Falvey. I leave Peacocke's arguments until second last, after I have responded to Gertler's arguments. This is because Peacocke's work shows us not only that pre-existing beliefs can involve transparency (as I have just argued above) but also that we can recall them immediately (without using evidence or inference).

Falvey

Falvey (2000) shows us how a remembered belief relates to its paired current belief. He argues that when the role of memory in pre-existing beliefs is properly understood, we can see that Evans' point about transparency holds with respect to standing beliefs. Memory of a pre-existing belief, Falvey argues, does not play a justificatory role relative to a judgment. It does not provide reasons or evidence for my current belief that I believe that *p*. Instead, it *presents as true* the content of the belief that *p* previously endorsed. It preserves my previous judgments and the justificatory relations among them,

[2] I have argued this way in Chapter 1, Section 1.

making them available for re-endorsement or reconsideration at later times (p. 81).

To simply cite my memory as a reason for thinking that I believe that *p* now would normally be odd, Falvey says. This is because to cite my memory about my belief that *p* is not yet to re-consider it by asking myself whether it is still true. But if I have previously considered the matter and endorsed *p* as true, these processes are available later where necessary, so that I can revisit the evidence for *p* to re-endorse it or change it.[3] My first 'I believe that *p*', when self-ascribed about an already existing, settled belief, is self-ascribed from the attributive perspective unless I can already remember what my reasons were for believing it in the first place. Otherwise, to remember it later, I must 'look into my mind', as Gertler claims. This, as pointed out above, is not an objection to Moran's transparency condition, which applies only to beliefs currently being considered by deliberation from the first-person, avowable perspective. Transparency does not apply to my report of my already existing belief because my report of this belief is probably attributive.[4] But there is a sense in which transparency was earlier involved in this situation because, as Falvey points out, a remembered belief presents that belief's content *as true* because transparency was involved in the subject's earlier deliberations that led to her believing it and to her knowing that she believed it in the first place. So the fact that we must usually first attribute to ourselves a remembered belief is not a problem for Moran. If I am uncertain that I still believe that *p*, I can now switch stances to re-deliberate whether I still believe that *p*. My re-endorsing of my earlier self-ascription at that point would occur from the deliberative, avowable

[3] Of course there may be the occasional situation where I am now unable to rework the evidence for my former belief (I can't concentrate for so long any more). In that case, we might take my earlier working through this evidence as sufficient grounds for my believing it now. It would depend on the context.

[4] See discussion of this in Chapter 1, Section 1.

stance, subject to the transparency condition. If I have been asked my date of birth, I might quickly remember that I still have my birth certificate and can just re-avow my date of birth, endorsing it straight away because I can justify it. If I am not sure, then I can re-consider. The key to seeing this involves the claim that the two stances are asymmetric. I can switch between these stances as often as I wish. I can cite my memory that *p* from the attributive stance (I can say that I used to believe that *p* and that as far as I know I still believe that *p*) and then switch to the deliberative stance to reconsider, via the transparency condition, whether and why I still believe that *p*.

Memory, Falvey explains in a footnote, 'is the voice of my former self "telling" me (perhaps with qualification)' (2000, p. 96, footnote 25) that what I believed earlier is still true, enabling me to re-endorse its content. But this is in itself not a re-endorsement. Re-endorsement must be based on re-deliberation.

I turn now to Gertler's arguments against pre-existing beliefs.

Gertler

Gertler argues that transparency cannot explain privileged access to our beliefs (Gertler 2011, pp. 125 to 145). One of her reasons is that transparency cannot be applied to pre-existing beliefs because to recall these we must 'look inwards'. She points out that even to answer Evans' own question, about whether there will be a third world war, I might remember that just this morning I told a friend I feared we were on the verge of such a war. Even to answer Evans' paradigm question, then, she claims, one need not necessarily use transparency (p. 125). As I have pointed out, transparency is not meant to apply to attributive attitudes on Moran's account. It is true that when we recall ourselves telling our friend our fear, we are telling him something for a reason or reasons that we can recall by considering our own minds, and thus not by using transparency. But we can then switch to the avowable stance to

reconsider (say) whether our fear is reasonable. We can do this also about the likelihood of another world war. At that point, once we are in the avowable stance, we are using transparency. We might decide that since country X did bomb Spain recently, and did walk out of the recent United Nations meeting, another war is imminent. Or we may have already considered just that morning why we believe we are on the verge of a third world war. In that case, we have already used transparency in forming that belief. Transparency is used when we are considering whether *p*, where *p* is something we are yet to make up our minds about. But for any belief that we have in the ordinary way, i.e., by being able to give, from the first-person perspective, justifiable reasons for our having it, we must have considered those reasons (among others) while we were making up our minds about whether *p* in the first place.[5] Transparency is thus 'built into' most of our ordinary beliefs. I say 'most' rather than 'all' because Gertler argues that we cannot do this with implicit dispositional beliefs, where the subject has not previously considered whether *p* and if asked whether *p* would answer immediately, without acquiring new evidence concerning whether *p*. I argue shortly, however, that transparency is also built into implicit dispositional beliefs. More generally, Gertler's point is not an objection against transparency if we assume, as Moran does, that the first-person perspective on self-knowledge is asymmetric with the third-person perspective.

Gertler's claim is that transparency cannot explain privileged access to our own beliefs (i.e., that we can know these immediately, without evidence or inference) because it cannot apply to pre-existing beliefs, to occurrent beliefs, to implicit dispositional beliefs, or to ordinary dispositional beliefs (2011, p.

[5] This obviously does not apply to beliefs that we have learned from another without ever have considered why we believe them. These we can only attribute to ourselves from the third-person perspective.

126). I consider each of these in turn.

Occurrent beliefs

An occurrent belief that *p* is a belief that we already have and are considering at the time we are asked if we believe it. But the transparency method allows us to consider only evidence that is available after the question is asked, not at the time at which it is asked. Transparency, Gertler argues, can only create new beliefs, not use existing beliefs. Suppose that by using transparency we *create* the belief that it is raining at t_2. We cannot also judge that it was raining a moment earlier, at t_1. We can remember that it was raining a moment earlier, although not, according to Gertler, via transparency (2011, p. 128–130). But why can we not adopt the avowable stance and thus look *outwards* to remember that it was raining at t_1? Why can we not apply transparency from the avowable perspective by using the evidence of the wet path that is before us right now (at t_2) that it was raining at t_1? Gertler can answer this objection, though. She can point out that although, when I am asked whether it is raining now, I can answer yes by using the evidence of the wet path, I will do this whether I remember or do not remember that it was raining at t_1. The question does not enable us to distinguish these two cases.

Gertler's objection, however, has already been answered by our discussion of Falvey's approach. Once we see that attributions do not require the use of transparency, but that we can switch stances as we please so as to answer Gertler's point, we can attribute to ourselves the claim that it was raining a moment earlier and then switch to the deliberative stance to avow that it is raining now, using the evidence of the wet path and of the rain falling on it.

Implicit dispositional beliefs

What Gertler calls implicit dispositional beliefs are beliefs we have never considered before but when asked about them can answer immediately

without acquiring any new evidence about them. Examples include being asked whether there are bicycles on the moon and whether bricks are edible.

Since such beliefs are dispositional rather than occurrent, Gertler's worry that transparency creates only new beliefs does not arise in these cases. Thus, transparency initially seems promising to the Cartesian as a method for answering questions about implicit dispositional beliefs. However, it turns out not to be, because implicit dispositional beliefs are beliefs that can be formed only by using pre-existing beliefs as evidence, and transparency, according to Gertler, cannot use pre-existing beliefs. So you can, of course, answer 'no' to both examples above, but not, on her account, via transparency. However, this ignores the vast amount of knowledge of the world in general that we all carry with us. We can answer 'no' so quickly to whether there are bicycles on the moon and to whether bricks are edible because we can quickly consider, from the avowable stance, that we know what bicycles and bricks are and what they are made of and what has been happening on the moon so far. That is, we consider current and past evidence, as well as our knowledge of bicycles and bricks. Even to consider whether we think there will be a third world war we must take into account our knowledge of what has been happening since the last world war.

Ordinary dispositional beliefs

Gertler's definition of an ordinary dispositional belief is as follows:

S dispositionally believes that p if:

- S has endorsed the content of p;
- S has stored this content in memory;
- S can readily recall that p.

To show that transparency cannot reveal ordinary dispositional beliefs, Gertler

uses the example of Nick (2011, p. 135), who was raised to believe that spilling salt will bring bad luck, which he can avoid only by immediately dropping a pinch of salt over his shoulder. Now, as an adult, Nick knows this is pure superstition but still feels compelled to drop some salt over his shoulder whenever he spills salt. When asked 'Do you believe spilling salt causes bad luck?' Nick will deny that he believes this. I suggest that he fears, rather than believes, that spilling salt causes bad luck. But why does Gertler deny that Nick can be using transparency? I suggest that Nick has formed his belief that spilling salt does not bring bad luck by realising that it is irrational and that he has reached this conclusion by considering how the world is. This is surely a case of transparency. Transparency is being used if your belief that *p* is transparent to the truth of *p*. *P* is true, in your eyes, if you have considered the external evidence about *p* and have arrived at this conclusion strictly on the basis of this evidence.

Gertler also discusses whether Nick has belief perseverance where spilling salt is concerned (2011, pp. 135–138). I suggest that this phenomenon is closely connected to Moran's concept of estrangement, in Nick's case an estrangement that comes and goes, its content sometimes available and sometimes unavailable to his conscious awareness. When Nick spills salt, since his real reason for dropping some salt over his shoulder is sometimes inaccessible to his conscious awareness, he cannot choose *not* to throw some salt over his shoulder at such times. On Moran's account, his avowal that he does not believe that spilling salt causes bad luck has no first-person authority at those times.

More importantly, there is a reason for Nick's sense of foreboding that is currently opaque to Nick (he is currently estranged from it). It may be, for example, that Nick is afraid that his deceased father will disapprove of him if he ignores the injunction to throw salt over his shoulder when he spills salt.

Fear of parental disapproval may remain active long after the parent has died. Once this reason becomes available to his conscious awareness, the foreboding may tend slowly to lose its force.

So although Gertler uses her discussion of Nick's problem with spilling salt as evidence against transparency, it seems to me to fit Moran's account very well; it shows a degree of psychological unfreedom in a case of estrangement. It also shows the subject's potential capacity to restore his authority over a previously estranged attitude by bringing it to conscious awareness. This exemplifies the structural relationship between authority and estrangement in Moran's account, discussed in Chapter 1.

I turn now to Peacocke (1998). Peacocke agrees with Falvey's claim that pre-existing beliefs are transparent; his work also links the issue of pre-existing beliefs to that of immediacy (to be explored in Section 2) by discussing how we get from knowing that we judge that *p* to knowing that we believe that *p*.

Peacocke

Peacocke (1998, pp. 71–72) uses the example of a subject who self-ascribes the belief that Dubcek was prime minister of Czechoslovakia when the Soviet Union invaded. The subject self-ascribes this belief for the reason that she has just then judged that Dubcek was prime minister at the time of the invasion. Peacocke says that we can take that statement at face value. He distinguishes three stages a thinker may pass through when asked, 'Who do you believe was prime minister there when the Soviet Union invaded?'

After reflection, i) she may have an apparent propositional memory that Dubcek was prime minister then. Another way of putting this would be to say that she believes he probably was but is unsure. Since she is, we may suppose, taking memory at face value in these circumstances and for this sort of subject matter, she then moves ii) to endorse the content of the apparent memory and

makes a judgment that Dubcek was prime minister then. This judgment makes it rational for her iii) to make a self-ascription of the belief that Dubcek was prime minister then (1998, p. 71). But if she is unsure at stage i), how can she be sure at stage ii)? I suggest that revisiting this matter can make her gradually clearer about what she is remembering, so that by step ii), she is confident about it.

Peacocke claims that to say that ii) is the thinker's reason for making the judgment in iii) is not to say that she infers iii) from the premise that in ii) she has made such a first-order judgment (1998, pp. 71–72). Using the example of a sensation for contrast, he argues that although an experience of pain can be a thinker's reason for judging that he is in pain, we should not construe this use of reasoning as a case of inference. If we did so construe it, this would make it impossible for us to explain the self-ascription of sensations, since we can hardly say that we have rationally concluded that we are in pain from the premise that we are in pain. It is the conscious pain itself, he argues, and not some alleged perception of it, that gives reason for the self-ascription. It is important for him to make this clear because he himself argues for an immediate, non-evidence-based account of self-knowledge, whereas perceptual accounts such as Fernandez's, to be discussed shortly, argue that self-knowledge is evidence-based.

Peacocke then argues that this case is not in competition with Evans' (1982) transparency procedure. When you search your memory to see if you know who was prime minister in Czechoslovakia when the Soviet Union invaded, he argues, you use Evans' procedure to answer the first-order question about who was prime minister then. Coming to self-ascribe a belief on the basis of the deliverances of stored information is a special case of use of Evans's procedure, he claims, rather than any kind of rival to it (Peacocke 1998, p. 73).

Peacocke continues that you would be justified in taking a shortcut in the

Dubcek example by moving straight from i), where you remember that Dubcek was prime minister then, to iii), where you self-ascribe your belief that he was, without using ii) at all: that is, without explicitly moving from remembering in i) to judging in ii) that he was prime minister then, before being justified in claiming to believe in iii) that he was. The short cut from i) to iii) in this example is justifiable, Peacocke claims, only where the thinker could have taken the longer route, as long as she could show that each transitional step in that longer route was made for the right sort of reason.

In this example Peacocke gives us a reason for claiming that pre-existing beliefs can be transparent by arguing that you can justifiably 'take a short cut' from remembering to believing without going via judging, as long as you can produce sufficient reason for your belief. As we will see, this is also a reason relevant to the discussion that follows shortly in Section 2, concerning how we can know immediately what our attitudes are. But before turning to this, we need to consider Reed's position.

Reed

Reed (2010) agrees with Gertler that pre-existing beliefs are not transparent.[6]

---

[6] Reed discusses accounts of self-knowledge that attempt, he says, to ground the special authority of self-knowledge in a constitutive relation between an agent's intentional states and her judgments about those intentional states. The accounts he discusses are those of Bilgrami (1998), Burge (1996), Moran (2001) and Wright (1996). Reed claims that the constitutive relation is non-epistemic (non-cognitive) because it is grounded not in a cognitive connection to an independently existing mental state but in 'the very nature of rational agency' (2010, p. 165). He argues that rational agency sometimes requires us to take an epistemic or cognitive stance towards our own mental states and that the constitutive account of self-knowledge needs to be amended to correct this deficiency.

As I have already argued, Moran's account allows for a purely epistemic stance; his attributive stance, which claims only that the subject *has* the belief, is epistemic if self-knowledge is substantive, a claim that was introduced in Chapter 1 and will be pursued in Chapter 4. The ordinary, practical stance that we use daily is Moran's avowable stance, which transcends empirical facts to *include* endorsement and commitment *as well as*, rather than *instead of*, being epistemic.

I have postponed discussion of Reed's position on this until now because his reason for making this claim is very different from Gertler's: Reed claims that Moran's deliberative stance is non-epistemic because his (Moran's) account of self-knowledge is constitutive. A constitutive account, Reed argues, is deflationist in a different way: that is, it claims that self-knowledge is non-substantive. Reed defines the constitutive thesis as the thesis that

> the special authority of a subject S's self-knowledge derives from the constitutive relation holding between S's first-order intentional state and S's second-order judgment about that first-order state. (2010, p. 165)

On this definition, since Moran's is a rational agency account and since rational agency is the heart of constitutivism as Reed defines this, Moran's avowable stance is non-epistemic. As already mentioned, Reed argues that the centrality of transparency to Moran's account is therefore a serious problem for it because there are many cases where transparency does not apply and yet where the self-knowledge involved is epistemic. I reject his claim by discussing his two examples of it. He claims that Moran's account must be amended to take into account situations where substantive but non-transparent pre-existing beliefs are required for an avowal to succeed.

But Moran's account of avowable self-knowledge, as I set it out in Chapter 1, is that it is epistemic-plus, rather than non-epistemic. If this is correct, Reed's objection from non-epistemicy fails. When you avow that *p*, you are also, in doing so, attributing to yourself the belief that *p*. Moran's deliberative stance *implies* attribution, since you can hardly avow that *p* if you do not even *have* the belief that *p*.

Reed's first argument for his position is that we become aware of our already formed belief about some topic by remembering it via a second-order epistemic process that underwrites the belief rather than by the deliberative process that first partially constituted the belief by finding it to be transparent

to its object. A remembered belief is constituted when it is first believed. Its later recall does not re-constitute it: it just brings to mind an already constituted belief (2010, pp 174–175). Thus a remembered belief, since it is believed at this point by what Reed calls a 'purely epistemic process', gives us self-knowledge that does not have the first-person authority of a belief newly constituted by deliberation. It thus falls outside the condition of active, rational agency.

But this, like Gertler's arguments, ignores the two stances and the subject's capacity to switch between them. Any successful, long-standing, pre-existing and newly remembered avowal, when it was first created, did meet the transparency condition and therefore did fall within Moran's account of active, rational agency. When it is recalled, however, the subject might not recall why he formed it. When this is the case, as Peacocke and Falvey have argued, the recalled belief is attributive only at this point. This is not a problem for Moran's transparency account, however, because although the subject might not need to re-deliberate right away in order to be sure of her belief, she can switch stances back and forth as she recalls her original reasoning and *deliberates* about whether she still believes that *p*. As we have seen above, both Falvey and Peacocke argue that a remembered belief *preserves truth*: that is, it presents *as true* the content of a belief previously endorsed. It preserves the subject's previous judgments, and the justificatory relations among them, making them available for re-endorsement at later times. So as long as transparency applies to remembered beliefs, the subject can re-endorse or change her previous belief at a later time. On Moran's account, and as Falvey and Peacocke argue, the subject can switch from the deliberative to the attributive stance and back again at will. She can remember what she just learned from switching to the empirical stance when she switches back again to the deliberative stance. She can hold both stances in her mind at the same time and weigh them up.

Reed's second objection is that we can sometimes have self-judgments of the form 'I believe that *p*' only by *failing* to meet the transparency condition: that is, we can be aware that we believe that *p* and can report this while being unable, at the time, to remember why we believe that *p*. In some situations it is rational for us to do this. Thus it is rational in some cases not to abide by the transparency condition (2010, pp. 176–178).

Reed uses the example of Penny the economist to support this claim. Penny, he tells us, began her career by writing several first-rate papers on taxation policy. Then her interests changed. Now, years later, a new colleague asks her about her views on taxation. Penny realises that she does not remember all the details of the views she laid out in her early papers. When this colleague asks her whether she now believes that *p*, this raises for her the question whether *p* (2010, p. 176). But it is rational, Reed argues, for Penny now to defer to her earlier judgment, given that no evidence against it has been produced in the meantime. She is not rationally required to re-evaluate it. She can simply decide that she still believes it. I agree. In the absence of counter-evidence there is nothing irrational about her doing this. It is rational for her to assume that if she took the trouble to do so she could re-evaluate whether *p* and decide the same way as before: that is, her currently presumed authority trades on her deliberative reflection, years earlier.

But of course we *can* make use of pre-existing beliefs in such situations. Penny *can* decide, instead, to re-evaluate her earlier opinions on taxation to see if she still believes them and still wants to avow them. She can do this because she can reconsider her earlier beliefs via their transparency or lack of transparency to the relevant facts as they are now. She can then avow, 'I believe that my earlier opinions are still correct (or are now incorrect)'. She is not attaining any new self-knowledge at this point; she is merely deciding to retain the belief she created earlier.

Summary and conclusion

The discussion about whether pre-existing beliefs are transparent begins with my own common-sense suggestion as to why they can be transparent, and then gives us Falvey's and Peacocke's positions on this, both of which support mine. Moving to Gertler, I object that her arguments that such beliefs are not transparent assume a quasi-Cartesian approach. I argue that her position *assumes* lack of transparency in pre-existing beliefs and then lists examples of this that follow only if her assumption is true.

Falvey and Peacocke, on the other hand, show us *why* pre-existing beliefs can be transparent. Both argue that remembered beliefs present their contents as true, so preserving the justificatory relations among these contents, and that they can be recalled in the attributive stance, re-deliberated in the avowable stance and either re-endorsed or not. We can *use* our minds to remember some part of the world as it is and as it was and to imagine or otherwise consider how it might be in the future. Peacocke and Falvey argue that we can normally move back and forth between the two stances at will. This shows us why Reed's argument against the transparency of pre-existing beliefs fails to show lack of epistemic substantivity—it is because he does not allow that the subject can switch between stances. I conclude that Moran's transparency account survives the objection that pre-existing beliefs are not transparent.

Peacocke's argument that the subject is justified in claiming to know that she believes that Dubcek was prime minister in Czechoslovakia when the Soviet Union invaded is also relevant to the discussion that now follows in Section 2, concerning how we can know immediately what our attitudes are.

Section 2: Our immediate knowledge of our attitudes

In this section I discuss the deflationary objection that our knowledge that we

believe that *p* is not immediate because it is inferred from the fact that we have judged that *p* on the basis of evidence. Even when we judge that *p* is true and know 'immediately' that we have so judged, it does not follow from this, the deflationist claims, that our knowledge that we believe that *p* is immediate. It does not explain *how* we get from knowing that we judge that *p* to knowing that we believe that *p* without inferring from judging to believing. Shah and Velleman (2005), Cassam (2010) and Fernandez (2003) have all argued that the step from knowing that we judge that *p* to knowing that we believe that *p* is evidence-based rather than immediate because it involves inference.[7]

Moran's (2012) answer to this objection is that it uses a concept of agency that is related to what he calls a 'production process': that is, a process of acting upon oneself to produce a belief without the need for using reasons. This is quite different from the concept of agency that Moran himself uses, to which the subject's reasons for her belief are central.[8] The 'production process' approach allows us to find out what we believe about something without considering the thing itself (p. 223), a claim that ignores the first-person perspective on self-knowledge by implying that at such times the subject does not use reasons in making her judgment about the object of her belief. In fact, the rejection of immediacy via the use of this approach goes hand in hand with the claim that the subject not only *need* not use reasons in forming her belief,

---

[7] Moran's explanation of this, already given in Chapter 1, is that in making reason-based judgments about the world, we are forming our own intentional states. Hence, it takes only ordinary linguistic/conceptual competence (such as the ability to know that believing something is just a matter of confidently judging it to be true) in order to move from judging that *p* is true to sincerely avowing 'I believe that *p*.' In fact, he argues that from the first-person point of view, judging that *p* is true and sincerely believing that *p* are conceptually rather than empirically related, so that to judge that *p* is true confidently or in a settled way, to use Moran's words (2001, p. 77), is already to believe that *p*.

[8] Both Moran (2012) and Boyle (2011) criticise the production process' idea. Boyle argues, for example (2011, abstract), that beliefs are acts of reason whose capacity for self-determination lies in their very nature, not in certain facts about how they originate. This is discussed in Chapter 4.

but also that she *must* not use reasons in doing so. Central to this method of explanation is that it involves the idea that I can ask myself whether I believe that *p* without considering *p* at all.

Moran argues that acting upon oneself externally to produce a belief in oneself is a passive rather than an active form of agency and therefore has third-person rather than first-person authority. Moreover, if belief formation did occur in that way, such acting upon oneself would produce only empirical, theoretical self-knowledge rather than the deliberative, avowable, first-person self-knowledge that transcends the empirical to include endorsement and commitment. The argument of this section is therefore that the objection to immediacy on the ground that deliberative, attitudinal self-knowledge is evidence-based fails.

To argue for this conclusion by using examples, I first discuss Shah and Velleman's (2005) presentation of this argument, followed by the accounts given by Cassam (2010) and Fernandez (2003). The arguments given by these philosophers are connected by the fact that they all, in various ways, attempt to eliminate the asymmetry between first-person and third-person self-knowledge by using a 'production process' account of agency.

Shah and Velleman: remembering might alter the remembered belief

Shah and Velleman argue that we can see that the appeal to rational agency (to a reasons-based account of self-knowledge) is misplaced by considering the difference between being asked 'Do you believe that *p*?' when *p* is something I already believe, and being asked 'Do you believe that *p*?' when *p* is something I have not yet considered. They say:

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

> applied to oneself for the empirical purpose of eliciting a response. One comes to know what one already thinks by seeing what one says in response to the question whether *p*. But the procedure requires one to refrain from any reasoning as to whether *p*, since that reasoning might alter the state of mind that one is trying to assay. Hence, asking oneself whether *p* must be a brute stimulus in this case rather than an invitation to reasoning. (2005, p. 16)

This approach requires that the person ask himself whether he believes that *p* without considering the question whether *p* (Moran 2012, p. 223). We are barred from *considering* whether *p* in a case where we already believe that *p* because if we do so, Shah and Velleman argue, we might alter the very proposition, *p*, that we already believe. So we must ask ourselves whether *p* and then wait until some answer appears spontaneously. Only in this way can we discover what we did believe at the time we were asked whether *p*.

I doubt that there is any problem here for a rational agency account of self-knowledge. Where we know in a settled way that we believe that *p*, we are unlikely to change our minds when asked 'Do you believe that *p*?' We usually just immediately recall our knowledge that we believe that *p*. Sometimes, though, when we are asked whether we believe that *p*, this might trigger our realisation that we have believed that *p* up until right now, but that now that the matter has been raised we are no longer sure that we still do. At other times, we might *mis*remember whether *p*.

None of this threatens Moran's account of immediacy because when we immediately recall our knowledge that *p* we are adopting the attributive stance. Transparency is relevant only to the avowable stance, where we are using reasons to decide whether *p*. If, however, we are unsure that we still do believe that *p*, we can *switch stances* to decide. We decide in reasoning mode.

It seems to me, therefore, that being asked whether *p* in cases where we already know that *p can* act as a brute stimulus. But this is not an either/or

situation. If it triggers in us the possibility that *p* might not be true after all, or reminds us of what our reasons were when we decided that *p*, it *also* acts as an invitation to reasoning. Such an invitation is not misplaced; it is accepted from the deliberative, avowable stance, to which transparency is relevant.

In the example that Shah and Velleman describe in their quotation above, if an answer does spontaneously occur to us and we assent to it, this is because we *know immediately* that we believe that *p*. If the question is 'Do you believe in God?', we might treat this as a 'brute' stimulus and say 'Yes' or 'No' immediately *because* we know it immediately, without inference—belief in God is usually something that remains settled over time. However, even here, the word 'brute' is misleading at best. One's previous belief in God may have been based on much deliberation at the time. It has a lot of personal history attached to it. But we may become unsure about whether we believe in God five seconds after saying that we do. In these cases, the question, 'Do you believe that *p*?' stimulates our reasoning capacity and we may discover either that we still believe that *p* or that we do not or that we are unsure. The act of reconsidering whether *p* does *not* necessarily alter the belief that *p* as it was before we were asked the question. We can simply *remember* that up until five seconds ago we believed that *p*. We may end up being convinced that we believe that *p*, and for the same reasons, or that we were wrong so to believe. Our spontaneous reply, 'Yes, I believe that *p*', is always subject to further consideration at a later date.

I conclude that, so far, Shah and Velleman's arguments fail to show that our attitudinal self-knowledge is based on evidence rather than being immediate.[9]

[9] This question is pursued in the next chapter, where I argue that Boyle's reflectivism refutes the claim that the step from judging that *p* to believing that *p* is inferred from evidence.

The question about how our knowledge of our own beliefs can be immediate is put by Moran himself as follows: how can the facts about *p* (where *p* is some situation in the world, such as that it is raining) have anything to do with, let alone constitute, my *knowledge* of what I believe about *p*? His answer is that I can know what I believe about *p* (that it is raining) if my reflections on whether *p* (on whether it is raining) have determined that belief. After all, if my reflections about whether it is raining have nothing to do with whether it is raining, whatever would**?**

He spells this out:

> One of the challenges to the Transparency claim can be put in the following way: what right have I to think that my reflection on the reasons in favour 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 favour 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 favour of rain provided an answer to the question of what my belief about the 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 the right sort of 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 matter? I don't normally think that my assessment of the reasons in favour of *p* might have nothing to do with what my actual belief about *p* is, and it's hard to imagine what my thinking would be like if I did normally take this to be an open question. And if I did think that my actual belief about the rain might be left quite untouched by my reflections on the weather-related reasons, what do I imagine could possibly close this gap for me? (2003, pp. 405–406)

In this, Moran asks: what does my belief about the rain have to do with whether

that belief is true? His answer gives us the commonsense position that it is *because* I can consider whether *p* (whether it is raining), by assessing the evidence for and against *p*, that I can claim that the belief that *p* that I form in doing so is *mine*, in the sense that I have formed it myself by using my own reasoning capacity about it. It is for *this* reason that my reflections on whether *p* can provide an answer to the question of what *my belief* about *p* is. It follows, therefore, more generally that my reflections on any event in the world can provide an answer to the question of what *my belief* about that event is.

Let us see now if Cassam's objection against immediacy, based on reliabilism, can refute Moran's position as just stated.

Cassam: we infer from judging to believing

Cassam's argument concerns one step in the reflecting process—the step from judging that *p* to believing that *p*. He argues instead that the conclusion of reflection on the reasons in favour of a proposition *p* is a judgment and that 'If I know that I judge that *p*, and am entitled to assume that my judgments normally determine my beliefs, then I can know or conclude that I believe that *p*' (2010, p. 88).

But this is not what we actually do when we form a belief about *p*. The fact that my judgments normally determine my beliefs does not imply that any particular judgment that *p* should or does determine that I believe that *p* in that case.

The problem for Moran's transparency account, however, according to Cassam, is that judging that *p* is based on evidence. Let us agree that judging that *p* is based on evidence. Cassam then argues that there is a sense in which my *belief that I believe* that *p* is also based on evidence, rather than being immediate. This, Cassam says, is because

> my judging that *p* is neither identical with nor entails that I believe that *p*. However, my judging that *p* normally leads (in the case in which I don't already believe that *p*) to my forming the belief that *p*, so the fact that I judge that *p* raises the probability that I believe that *p*. It makes it likely that I believe that *p* and is, in this sense, a reliable sign that I believe that *p*. But this is just what it is for one thing to be evidence for another. (2010, p. 89)

A sign of something, Cassam argues here, is evidence for that thing. In the sense that judging that *p* makes it more likely that I believe that *p*, judging that *p* is a reliable sign that I believe that *p*. But this means that judging that *p* is evidence that *p*. And this, he argues, implies not only that my knowledge that I *judge* that *p* is evidence-based, not immediate, but also that my *belief* that I believe that *p*, since it is based on the evidence that I have already judged that *p*, is also evidence-based rather than immediate.

This argument would normally be taken as asking us to suppose that having considered the evidence and judged that it is raining, we need to consider whether this judgment itself is *further* evidence that it is raining, in order to conclude that we believe it is raining. But we do not normally require more evidence about whether it is raining in order to move from judging that it is raining to believing that it is raining. We *make up our own mind* about whether we believe that it is raining *based on* the evidence for this that we have *already* used in order to *judge* that it is raining. The judging is not in itself further evidence. Further evidence would be (say) that someone is hosing down the roof above the window. If we think *that* might be happening, we might suspend our belief that it is raining, but unless there is reason to suppose something like this, we *already believe* that it is raining. We do not go *from* judging *to* 'believing'; in judging, we are already believing. This point is made by Boyle in his 'Reflectivism', and set out in Chapter 4. But in a commonsense way, as long as we consciously know that we are deliberating about whether *p as we do* this deliberating, we also know our answer *when we make it*. When the evidence is

believable and acceptable, we believe that *p* in making the judgment that *p*. The difference between judging and believing that *p* is often just that the judgment becomes a belief once it is settled.

Cassam does not mean his argument to be taken in the normal, commonsense way. He uses it, instead, in a production-process way. My judging that *p*, since it is something that I do, is evidence about me. It might be a reliable sign, and thus in this sense evidence, according to the production process position, that I am likely also to believe that *p*. But it is not evidence about *p*; it is evidence about *me*. As such, it is vulnerable to the same objections as those we have already made against Shah and Velleman. It says that to know whether I believe that *p*, I must use evidence about myself, just as I would about someone else, and must not use evidence about *p*.

Now clearly, if I must use evidence about myself, the knowledge that I conclude I have (the knowledge that I believe that *p*) will not be immediate: it will be evidence-based. For example, I might say to myself, 'Well, I judge that *p*, and judging that *p* is a reliable sign that I believe that *p*; therefore, I believe that *p*.' But we do not carry out any such procedure. Instead, we simply realise that as we are confident about our judgment that *p*, we now believe that *p*.

I conclude that Shah and Velleman's and Cassam's objections to Moran's claim that self-knowledge can be immediate all fail for the same reason: they use an approach that is insensitive to the asymmetry between the first-person perspective of the subject as active agent and the third-person, empirical perspective. The fact that this kind of argumentation fails is evidence that the first-person perspective is intrinsically asymmetric with the third.

The puzzling nature of the deflationary arguments we are discussing is intuitive support not only for the claim argued for in Section 1, that we can look at the world to recall pre-existing beliefs but also provides support for Moran's claim

that we need both stances, and they are irreducibly asymmetric.

Cassam falls back on a 'sub-personal monitoring box' solution: such a box can by-pass judging that *p* and take us straight from believing that *p* to knowing that I believe that *p* without observation, evidence or inference from judging to believing. In this way, he argues, a sub-personal solution can explain our self-knowledge that *p* by reference to something that the subject, in a sense, 'does' (p. 93). But this would imply that we do not need even conviction to move from judgment to belief. Clearly, we do need conviction, and thus conscious awareness. However, we can use Peacocke's suggestion[10] that we can omit the step of judging that *p*. Peacocke's suggestion involves being able to fill in this step with satisfactory reasons that the subject is consciously aware of. Cassam's explanation does not give us satisfactory reasons. If we know that we believe that *p* only because this knowledge has popped into our minds from a sub-personal belief box, this does not in itself tell us what our reasons are for believing that *p*. Since we can consciously make up our own minds what our reasons are, we do not need a sub-personal belief box.

Reliability and probability cannot secure self-knowledge in any particular case. This fact bedevils Fernandez's account as well. I turn to Fernandez now.

Fernandez: a deflationary account based on perception

Though he accepts a version of the transparency approach to self-knowledge, Fernandez claims that 'one may have the very same grounds for both a given belief that *p* and a higher-order belief about this belief' and that his account of self-knowledge requires only those concepts that we already use to account

[10] See this chapter, Section 1, for Peacocke's argument for this claim.

for perceptual knowledge (2003, p. 352). This assumes what Boyle calls the Uniformity Assumption, that there is only one kind of self-knowledge, not two kinds, as the Kantian tradition would have it.[11] Fernandez also claims that his account, if viable, would have the advantage of constituting a naturalising account of privileged access that does not posit any 'mysterious faculty of introspection or 'inner perception" mechanism' (p. 352). I would point out that neither Moran's nor my own account of immediate access, for example, posits any such faculty.

However, if the Uniformity Assumption is wrong and there are two kinds of self-knowledge, as Moran and Boyle both argue, avoiding 'inner perception' does not imply accepting that our knowledge of our beliefs is symmetric with our knowledge of our perceptions and sensations. Fernandez, like the other deflationary writers discussed in this chapter, uses reliabilism to argue that evidence that *p* will generally lead to a belief that *p*. On this reliabilist view, if you have evidence that *p* you are *warranted* in self-attributing the belief that *p* because believing that one believes the propositions that are supported by one's evidence for them is generally a reliable belief-forming process. On this view, a subject need not be able to produce any reasons for her belief; immediacy is explained as reliability of evidence.

From the point of view of a commonsense folk psychology this is not an adequate justification of a subject's self-knowledge because it may lead you to claim justification for knowing or believing something for which you have no normative warrant. Support for this claim comes from Falvey (2000) and Zimmerman (2004). I begin with Falvey.

Falvey points out some important differences between a perceptual and a

[11] I gave Boyle's argument against the uniformity assumption in Chapter 1, Section 2.

rational agency account of self-knowledge:

> The broad, perceptual theorist of self-knowledge … must think he can parlay the mere statistical fact that perceptual beliefs usually co-vary with the states of affairs that would make them true into a notion of warrant for these beliefs. This is to ignore all the normative features that justify locating perceptual judgments within the … 'space of reasons'. If statistical reliability is all there is to perceptual knowledge, then one really should attribute such knowledge to thermometers and weather-forecasting bunions. (2000, p. 86)

Ignoring all the normative features of a perceptual judgment clearly makes for an inadequate description of a self-ascription from the first-person perspective. Falvey also points out that a perceptual model of self-knowledge 'represents the subject as a passive spectator of the belief-forming processes within him, and reduces his self-ascriptions to something like the scripted pronouncements of a government spokesperson' (p. 87). This sounds very much like the deflationary production process method. He then points out that the perceptual theorist's approach to self-knowledge 'severs the warrant for the avowal from its source in the subject's authority, as a rational being, to make up his own mind on the question of what belief is best supported by the evidence.

These arguments of Falvey's support Moran's position on deflationary objections to the immediacy of self-knowledge. They also help us to understand why the objections being discussed here seem so puzzling. It is simply because, as Falvey points out, we are not passive spectators, pronouncing ideas scripted by a government spokesperson. Deflationism seems strange because it is strange: it gets things back to front.

Zimmerman adds to the criticisms of a perceptual account of our knowledge of our attitudes, such as that offered by Fernandez, by arguing against Fernandez's claims concerning the transparency of beliefs. He claims that the

conditions Fernandez describes are either unnecessary or redundant:

> despite their appeal, the conditions for introspective justification that Fernandez describes are not necessary for introspective justification, and … they could only be sufficient at the cost of being redundant. (2004, p. 436)

The example Zimmerman uses first to support this claim is of Mary, who has excellent evidence that the biological differences between species can be explained by natural selection but remains unconvinced. She may sincerely report that she is certain that the theory of evolution is true but when asked if it is true will refuse to give an affirmative answer. She does not go so far as to affirm a Moorean paradox[12] by asserting '*p*, but I do not believe that *p*', but she comes close enough to doing this to put her rationality in question.

Zimmerman points out that there is an important difference between a perceptual example and a belief example such as this one. Mary is in a state of theoretical psychological certainty that evolution is true because of the excellent evidence for it. On the evidence she has, she would be justified in avowing that evolution is true. But when asked, she cannot avow that it is true. Zimmerman argues that Mary's sincere report, that she is certain that the theory of evolution is true, might *seem* to be similar to the perceptual mistake of misperceiving a fake apple as a real apple. But this perceptual mistake might not even have been a mistake, since there can be such things as fake apples, and even if it is a mistake it could well be a justifiable mistake. However, Mary's report of her theoretical belief that evolution is true, followed by her refusal to assert that it is true when asked, involves reasoning that is too close to being paradoxical to be justifiable. (In my view, we can explain this by saying that Mary's attitude towards evolution is conflicted. Because of this, she cannot use her theoretical knowledge that evolution is true in a practical way.)

---

[12] See Moran 2001, pp. 69–77 for a discussion of Moore's Paradox.

So if a subject can refuse to assert that *p* and yet sincerely report that *p* because of the excellent evidence for it, Fernandez's conditions for justification of an avowal of a belief are insufficient; the subject must always *really* (sincerely) believe that *p* and be prepared sincerely to avow that *p*.

But now, Zimmerman continues, suppose that *whenever* we know subjectively that *p* (via evidence) this knowledge justifies our believing that *p* without any subjective uncertainty. In this case, since our second-order beliefs are based on our first-order beliefs anyway (being always taken as true of them), there is no need to argue that the second-order beliefs are grounded in the evidence that supports the first-order beliefs; this is now redundant.

Also, Zimmerman goes on to say, it is misleading to say that 'the subjective justification for the belief that *p* justifies the belief that one believes that *p*' (2004, p. 437). He gives the following example to support this second contention: you feel pain in your foot. You look down and discover you are standing barefoot on a piece of broken glass. But you felt the pain before you discovered the glass. Your avowal 'My foot hurts' does not depend on your knowledge of the evidence of the broken glass to justify it subjectively. You know your foot hurts without needing any evidence at all.

Falvey's discussion (2000, p. 86) of the failure of the perceptual model to accommodate normative considerations also supports Zimmerman's arguments. If this model were true, our 'avowals' would resemble government scripts mindlessly recited.

The conclusion of the foregoing discussion of deflationary objections to Moran's transparency account has been argued for by Moran himself. Moran points out that since a distinctive feature of first-person discourse is 'that a person can answer a question about her own belief by addressing herself to the corresponding question about the topic of that very belief' (p. 212), then if

transparency is ever legitimate 'It must represent a systematic difference between relations to oneself and relations to others' (2012, p. 213).

This means that a transparency account of self-knowledge must be an account *from the first-person position*, since the identity of the person answering the question and the identity of the person whose state of mind is being inquired into (Do you believe that *p*?) must be the same. If I use its transparency to the truth of *p* as I see it in order to determine what my belief is, it must be I myself who determines what my belief is. I cannot use the fact that Jones believes that *p* in order to use *p*'s transparency to its truth as Jones sees it in order to decide what *I* believe about *p*. If I believe that *p* only because Jones believes that *p*, I do not do so by using transparency. I know only that I believe that *p because* Jones believes that *p*. This is to focus on Jones, not on the object of *p*. It is not evidence about the object of *p* that I can use in making up my own mind about whether *p*. Thus the belief does not conform to Moran's transparency condition; the subject's self-ascription of it is attributive, not avowable. The conclusion is that the first-person position cannot be rendered symmetric with the third-person position; an account of self-knowledge based on the subject's own reasons must be an account from the first-person position.

Here is a commonsense position as to how I know that I believe that *p* when I know that I judge that *p* and take what I believe to be determined by what I judge. It begins with Peacocke's claim, discussed in Section 1 of this chapter, that when I am asked whether *p*, I can consider the evidence for and against *p* with full, conscious awareness that I am doing so. If you ask me what I am doing, I can say, 'I am considering whether *p*'. So I can already know that I am considering whether *p* as I continue to consider whether *p*. Then, wholly on the basis of this considered evidence, and conforming to the transparency condition, I judge that *p*. Again, I do this in full conscious awareness of what I am doing. I therefore now know that I judge that *p*: I can avow, 'I judge that *p*'.

So far, so good.

Now we need to add only that I judge that *p* not only in full conscious awareness of what I am doing but also *with conviction*. In this case, there is nothing further I need add to be able to avow that I believe that *p*. I can endorse *p* as true and can commit to behaving responsibly towards the object of my belief. I can say, 'I know that I believe that *p*'. Standing in the pouring rain, wet to the skin, I can avow, 'It is raining.' I do not need to consider the statistical probability that I believe that it is raining because I have just judged that it is raining. My conclusion does not depend on any statistical probability, however reliable.

Adding 'with conviction' does not insert an inferential step between my judging that *p* and my believing that *p*. To say that I judge that *p* with conviction is to say only *how* I judge. I do not judge lightly or casually or doubtfully, I judge with conviction. The addition of 'with conviction' is adverbial: it does not denote a separate mental state. It can occur because I am totally focused on what I am doing; I can say, 'I am convinced that *p* is true' and thus also, 'I know that I believe that *p*', because I know what I am doing *as I do it*. In fact, I have known all along what I was doing. We can actually say (to anticipate Boyle's discussion of reflectivism in Chapter 4) that I judge *knowingly* that I believe that *p*.

This does not imply that I always know what I believe. I may judge that *p* with conviction, then immediately forget what I have just judged and be unable to remember it ever after. It does imply, though, that there are many cases where I can know, immediately, without further evidence or inference, that I believe that *p*, once I have already judged that *p* by using evidence. However, this is only a commonsense position. It needs to be spelled out philosophically: this is what Boyle does in Chapter 4.

Summary and conclusion

In this chapter I have discussed two major objections to Moran's transparency condition: that pre-existing beliefs are not transparent and that transparent self-knowledge is not immediate. Both objections aim to show that first-person authority is symmetric with third-person authority and therefore that first-person self-knowledge is symmetric with third-person self-knowledge. I have argued here that these deflationist objections to Moran's transparency condition do not succeed, and if the Moran/Boyle argument for the distinction between active and passive agency[13] is viable, this deflationist position cannot be sustained more generally either. The avowable is normative; it is substantively epistemic and it is *also* about endorsement and commitment. The transparency of the avower's attitude to the truth or choice-worthiness of *p* is an essential ingredient in the structure of Moran's account. Moran can claim that a subject can know that she believes that *p* because by using the phenomenon of transparency she can form the belief that *p* in a reasons-based way that is answerable to the external facts. The avowable stance cannot, therefore, be rendered symmetric with the non-avowable, non-reasons-based stance.

I have defended Moran's position against deflationary objections in this and the previous chapter because the structural relation that Moran spells out in *Authority and Estrangement* between first-person authority and estrangement depends on asymmetry. Deflationist accounts cannot explain repression, estrangement or the anomalies because they cannot conceptualise any relation either between estrangement and first-person authority or between first- and third-person self-ascriptions. These relations, as Moran lays them out, place estrangement within an account of self-knowledge in a way that can

[13] This position is set out in Chapter 1, Section 2.

explain how it motivates the anomalies. By placing it in opposition to first-person authority, Moran can explain passive agency in terms of it, as, for example, in rejecting Cassam's argument against the immediacy of transparent self-knowledge.

In Chapter 4 I argue that Boyle's reflectivism transforms the conceptual landscape of philosophies of self-knowledge. In Chapters 5 and 6 I discuss estrangement and the anomalies with respect to self-deception in Chapter 5 and akrasia in Chapter 6.

Chapter 4: How can we know and not know what we believe?

Introduction

Moran argues that the question of how we can know what we believe is conceptually connected with its opposite, that of how we sometimes *cannot* know what we believe. I have argued, using Moran's account as my source, that the answer to this latter question lies in the phenomenon of estrangement. Of course, because of brute factors such as mental confusion, we sometimes cannot know our own minds *in fact*: that is, empirically. However, the only way we cannot know our own minds *in principle,* if Moran is right*,* is because we suffer a relevant estranged attitude.

In this chapter, I continue to support Moran's claim about estrangement via Boyle's concept of reflectivism. Boyle (2011) argues that what is usually described as the first-order state of my knowing that I believe that *p* is actually the more complex first-order state of my *knowingly* believing that *p*. I set out his account of this in Section 1 and discuss some consequences of it in Section 2.

Boyle's argument is based on the Kantian distinction already discussed (in Chapter 1, Section 2) between active agency and passive agency, a distinction that I applied in Chapter 3 to deflationary objections to transparency. In this chapter I apply it again, this time to the problem of how we know what we believe.[1] If an avowal is of the first-order state of *knowingly believing* that *p*, we do not need to say that our knowledge of it, once acquired, is evidence-

[1] This Kantian distinction is also useful in Chapter 6.

based. To show that it does not imply this, we need to show that both our judging and our believing that *p* on the basis of evidence are normally performed knowingly.

Reflectivism is important to this thesis firstly because it brings us to estrangement. If it is viable, it shows us how we can have tacit knowledge of an attitude that we have but that is inaccessible to our conscious awareness. But if we can be estranged from our self-knowledge, it might be this estranged attitude that causes the anomalies by motivating them. If estranged attitudes can motivate the anomalies, then, secondly, they must be substantive.[2] How could something that is not an independently existing attitude motivate the subject to form a certain belief or to act in a certain way? Here is a reason for claiming that self-knowledge is substantive. Only if estranged beliefs can become active at least sometimes, can they be said to motivate the anomalies.

Section 1: Boyle's arguments for reflectivism

As Boyle expounds it, the theory of reflectivism holds that our *explicit knowledge that* we believe that *p* (when we have it) is an essential aspect of our belief that *p*. It uses Moran's claim that such knowledge is based on our capacity to make up our minds about what we believe *by reflection*. Thus reflectivism is an account of how we can know our own attitudes based on reflection.

To introduce Boyle's thinking about reflectivism, I begin with his answers to Byrne (2011). Though Byrne is himself an advocate of a transparency view (unlike some other opponents of reflectivism, like Gertler), he objects to the

[2] Substantive' means 'having a separate and independent existence, not merely inferential or implicit', or 'the genuine "detection" of some independent psychological fact' (Moran 2001, p. 13).

way Moran develops his transparency account. I begin with Boyle's discussion of Byrne's account of self-knowledge.

Byrne tells us (Boyle 2011, pp. 224–225) that his (Byrne's) account of self-knowledge is economical: it requires no power of self-determination and thus no active agency; it requires nothing more than the ordinary human capacity to draw inferences where both perceptions and attitudes are involved.[3] It is essentially an empirical 'self-knowledge as production' account. Boyle's procedure is to show that Byrne's explanation of self-knowledge cannot be right. He does this by contrasting the idea that self-knowledge is inferential[4] with his own idea that it is reflective; so his argument for reflectivism[5] opposes, at the same time, an inferential account such as Byrne's.

Boyle first points out that we are not reliably disposed to believe that some external fact is true just because it is true. When we consider in a commonsense way whether (say) sea levels are rising, it becomes apparent that we would not be investigating this particular question unless we had a reason for doing so. Perhaps someone has asked us this question; perhaps we have become interested in sea levels because we own a property by the sea; perhaps we have to pass an examination next month on climate change. But for whatever reason, we have engaged with this question and we are paying attention to it. The situation is not, therefore, that if sea levels are rising we are reliably

---

[3] In his inferential account, Byrne rejects the Kantian idea that Moran and Boyle both argue for, that there are two different kinds of self-knowledge of our intentional attitudes.

[4] Boyle describes an inferential account of transparency as one where there is a non-accidental transition between belief contents, where the reasonableness of the transition is open to assessment (2011, p. 227).

[5] Boyle describes a reflectivist account as one where the step from believing *p* to reflectively judging (that is, consciously thinking to herself: 'I believe *p*') is explicitly to acknowledge a *condition* of which one is already tacitly aware (2011, p. 227).

disposed to believe this. We are not disposed to believe that *p* just because *p* is true. Millions of statements about worldly events are true without our believing or being disposed to believe that they are true. We will be likely to believe that they are true only when we have some reason for believing this. But for this to be the case, the question whether *p* must first have engaged our attention and we must feel that there *might* be a good reason for us to form a belief about it. Otherwise, we would not be reliably disposed to form a sustainable belief about any worldly fact.[6]

This is a very different situation from that where when you believe that *p* you do so automatically, where just the fact that *p* somehow leads you to believe that *p*, as the 'production as process' argument discussed in Chapter 3 claims. How could it do this? The situation is not, for example, that there is the world and there is you and there is an automatic connection between the two that causes you to know what there is in the world by some sort of stimulus-response. You are not omniscient. You have to *make* this connection, *for some reason*, before you can even begin to consider whether *p*. Boyle draws this conclusion by arguing against Byrne's doxastic schema as follows (Boyle 2011, pp. 230–232):

Byrne's doxastic schema

In Byrne's schema:

- *p*.
- I believe *p*.

Or, using the sea levels example:

- Sea levels are rising.

[6] Boyle argues this way in 2011, p. 231.

- I believe sea levels are rising.

Byrne's schema says that from a proposition *p* about the world, such as 'sea levels are rising', we can infer to a proposition about what I believe, such as 'I believe sea levels are rising.' Boyle's demonstration of the invalidity of this inference is as follows:

> To believe that I believe *p* is to hold it true that I believe *p*. So to believe that I believe sea levels are rising is to hold it true that I believe sea levels are rising.
>
> Suppose, now, that I ask myself on what grounds I hold it true that I believe sea levels are rising. The answer 'because sea levels *are* rising' is irrelevant to this question, because even if sea levels are rising, this, by itself, has no tendency to show that I believe that sea levels are rising. There are many true propositions about the world that I do not believe. It is also no use saying that I believe sea levels are rising on the ground that I hold it true that I believe sea levels are rising. This ground is redundant: I already know, in this case, that I believe sea levels are rising and so I do not need to infer it from elsewhere (p. 230).

Other reasons that Boyle gives against the idea that inference deposits beliefs in my mind in a merely stimulus-response way are that I must be cognisant of such a transaction, that I must take it that there is an intelligible relation between the terms of the worldly fact and of the belief and that I must be able to see a reason for making such an inference (p. 231).

The subject is knowingly involved

These reasons imply that the subject is knowingly involved in the transaction. The explanation of how we move between propositions about the world and propositions about what we believe must therefore be two-way: that is, the relevant field of inquiry must include an aspect of the believer: her capacity to know what she is perceiving or thinking as she perceives or thinks it. Moreover, Boyle argues, she observes the world under a specific mode of presentation,

such that the form of her judgment implicitly bears on the nature of her mental state:

> she does not arrive at knowledge of one realm of facts by inference from another, epistemically independent realm of facts (p. 233).

Epistemically speaking, we and our world inhabit the same field of inquiry; the interactions between us and our world are therefore two-way. It is this that enables the believer to engage with the world in a way that makes it possible for her to know what she is doing as she does so.

This implies that the field of inquiry into how I know what I believe includes both me and the world. I and the world are both *in* this field. There is one boundary around both of us. My thoughts, perceptions and acts of reasoning interact with the relevant external events within it to enable me to answer the question whether *p*. So it is hardly surprising that these thoughts, including my capacity to know what I am doing and knowingly to form beliefs about this on the basis of evidence, may also already be present when I form my answer.

This is not, therefore, a situation where we *infer* to what we know (our own psychological state) from facts about the world. To do this, we would first have to identify what it is that we want to know about. But in the normal case, in doing *this*, we are already knowingly connecting with this thing as something that falls within our capacity to form beliefs about. We are already drawing a boundary around us that includes us both. The basic element in knowing our own attitudes is not making inferences, although we may well do this as we go: the basic element is that in forming a belief we are acting for a reason. We apply our capacity to form beliefs about the world *for reasons,* by acting as rational agents, and we do this knowingly.

If Boyle's argument against Byrne's inferential account of self-knowledge is viable, reliabilism cannot explain *any* cases of self-knowledge, with or without

first-person authority, because reliabilism is a one-way explanation.[7] One might object to the idea that I form beliefs for reasons by pointing out that often I can know that I believe that *p* without being able to give any reason at all for this belief. But normative considerations will always apply. In the normal case we are expected to *have* a reason for the belief we have formed, to be able to state this reason when appropriate and sometimes to show some sort of reasonable behavioural commitment to its truth.

Tacit belief and reflection

Before discussing reflectivism further, we need to be clear exactly what Boyle means by 'tacit'.

For Moran, tacit belief is belief that we take for granted but have never reflected on explicitly (2001, p. 29). 'Tacit knowledge' is typically used in two ways: to refer to claims we have previously endorsed but which we are not currently entertaining, and to refer to dispositional knowledge we have never previously endorsed. In Moran's usage, it refers to dispositional knowledge: we tacitly know those beliefs that we have never, as yet, made explicit via reflection, although we have taken them for granted.

Boyle uses Moran's way of using 'tacit'. As Boyle explains it (2011 p. 228), a subject who believes *p* has some tacit awareness of believing that *p* before he becomes consciously aware that he believes that *p*. Let us suppose also that all the evidence is in, so to speak, concerning whether *p*, and that this evidence, on balance, implies that *p* is true. In that situation, Boyle argues, the subject normally knows tacitly that he believes that *p*. I suggest he knows this in the

[7] It is one-way even though, on Shoemaker's account (2009, p. 36ff), it is not causal because the relation between the two relevant states is constitutivist and metaphysical rather than epistemological and contingent. Its constitutive nature (if it has one) does not affect its 'one-wayness'.

same way that the person who is asked the way to the post office knows tacitly that you have to turn left at the next corner but has so far not become explicitly aware of this fact because he has never needed to become aware of it until now. So how does this person get from tacitly knowing to explicitly knowing that you have to turn left at the next corner to get to the post office? He gets there, Boyle says, by reflectively judging: that is, by consciously thinking to himself: '*p*' (in my example, '*p*' is 'You turn left at the next corner').

If, once all the evidence is in, reflection is all you need in order to judge (on the basis of evidence) and so to believe that *p*, then you do not need to infer from one psychological state to another (that is, first-order reflection, not second-order or metacognitve reflection, is all that is needed to make one's knowledge explicit). Boyle underlines this when he says that believing *p* and knowing oneself to believe *p* are not two different psychological states: they are two aspects of one cognitive state, that of knowingly believing that *p*, either consciously or unconsciously. If this is viable, it changes the landscape of the debate about substantivism because the arguments for the default position and for non-substantivism depend on considerations about the relation between a first-order belief about the world and its paired, second-order state. If there is only one state to consider, those arguments are irrelevant.

'Tacit belief' in the Boyle/Moran sense is also what Shoemaker (2009, p. 40) calls a latent belief. For Shoemaker, a latent belief normally becomes available—accessible to the subject's conscious awareness of it—only when it is relevant to her current concerns or when the question of its truth is raised.

'Tacit knowledge' can also be used to mean an ordinary standing belief, similar to what Gertler (2011, p. 134), describes as an ordinary dispositional belief, where the subject has already consciously endorsed this belief's content and stored it in memory, from where he can later recall it when required. But the man who is asked the way to the post office holds a tacit belief that he has

never made explicit to himself until now. And yet he knows what he is doing as he gives directions to his questioner. He does not need to consider whether his questioner should go by route A or route B. He already knows that route A is better. This shows us that he does know how to get to the post office, once he reflects on his knowledge of the streets between himself and the post office. So it shows us that tacit knowledge that has not previously been made explicit to the subject's conscious awareness does exist.

Gertler (2011, p. 131) refers to such tacit beliefs as 'implicit dispositional beliefs', where the subject has not previously considered whether *p* and yet can quickly assent to *p* if asked, without acquiring any new evidence concerning it. For example, when asked whether there are bicycles on the moon or whether bricks are edible, she can immediately say 'No', having never previously considered either proposition. Shoemaker (2009, p. 40, n. 9) refers to these as tacit beliefs.

Boyle is saying that we can form the belief that *p* without being consciously aware that we have done so, and that in some situations this can count as knowing tacitly that we believe *p*. On some occasions we make our belief explicit via (first-order) reflection and then store it in memory as a standing attitude, from where we can normally recall it to our conscious awareness as an occurrent episode when we need to. At other times, we just continue to know, tacitly, that *p*.

This conclusion transcends previous philosophical argumentation about how we make the step from believing that *p* to knowing that we believe that *p* (whether by inference or by reason). This is because the conclusion of Boyle's argument is that 'believing *p* and knowing oneself to believe *p are not two cognitive states*; they are aspects of one cognitive state—the state, as we might put it, of knowingly believing *p*' (p. 228).

This conclusion fits our knowledge that we know how to get from A to B without ever having consciously thought about it. The man asked the way to the post office knows how to get there without being consciously aware that he knows this. When asked, he simply draws on this previously tacit knowledge to give directions. He does not draw on any further evidence to support this already existing tacit knowledge.

Believing and knowing that one believes that p

With this conclusion, Boyle bypasses, with one step, the problem of how we know what we believe. If his claim is viable, we need not consider whether a belief and our knowledge of this belief are two distinct states or not. They are not two distinct states:

> The reflectivist rejects an explanatory demand that many theorists of self-knowledge accept. He denies that, in the normal, non-alienated case, being in a given mental state M and believing oneself to be in M are two distinct psychological conditions, and consequently denies that the task of a theory of self-knowledge is to explain how these conditions come to stand in a relation that makes the latter knowledge of the former. (p. 235) [8]

On the reflectivist approach, there is only one state involved in our knowledge of our own attitude. But this one state includes, as an adverbial aspect of it, the fact that we know it because it is the mental state of knowingly believing that *p*. It remains active for as long as the subject holds the belief to be true.

Section 2: Some consequences of reflectivism

If reflectivism is viable, some important consequences follow. The main

[8] Boyle claims that reflectivism applies only to some kinds of judgment sensitive attitudinal states and not in the same way to all of those kinds. Different accounts might be needed to differentiate fearing from wishful thinking, for example.

consequences are about the epistemic status of self-knowledge, its commonsense realism, and its ability to explain denial where estrangement and thus the anomalies are involved.

At first glance reflectivism seems to pose a problem for any account that claims to be epistemic-plus, because Boyle describes his reflectivism as non-epistemic. But it is non-epistemic only according to his own reflectivist approach, and only in a certain way: he denies that in the normal case there are two different states involved in self-knowledge. There is only one such state, he says: that of knowingly holding a certain attitude. Also, this 'one state' account is consistent with claiming that the mental state of knowingly believing that *p* is itself an epistemologically substantive attitudinal state: we can claim that attitudinal self-knowledge is epistemologically substantive as well as metaphysically constitutive. Boyle footnotes [9] that he avoids the label 'constitutive' because this label is sometimes taken to imply that judging oneself to believe *p* makes it the case that one believes *p*, and that is not his view. Boyle's idea is that 'for a rational creature, believing *p* just is being in a condition of actively holding *p* to be true' (p. 236), whether the subject has actively made up her mind about this or holds the belief without conscious reflection, such as when she looks out the window and forms the belief that it is raining. In this sense, he argues, our beliefs are normally examples of our capacity to make up our minds: 'They are all enduring actualisations of our power to evaluate propositions as true, in the light of such grounds as we deem relevant'.

Secondly, reflectivism is consistent with a commonsense realist account of self-knowledge from the normative, first-person perspective. It gives us a way of avoiding the causal sequence or 'production process' method by which Cassam

[9] Boyle 2011, pp. 228–229, n. 5.

(2010), among others, seeks to show how we acquire self-knowledge. It claims that we do not acquire it by any such method because we can already know what knowledge we are seeking as we begin to seek it. This explains why production process arguments seem puzzling and back to front. This is because they *are* back to front.

Boyle (2011, p. 233) argues for this claim, as it applies to Byrne's theory of our knowledge of our intentions, as follows: he says that Byrne proposes that I can infer 'I intend to *X*' from 'I will *X*', provided that I do not believe that I will *X* on the basis of good evidence. To bring out the oddity of this proposal, Boyle distinguishes two uses of 'will *X*': one where I already intend to *X*, and one where I might not already intend to *X*. If I already intend to *X*, then the phrase 'will *X*' can express my intention to *X*. But this use of 'will *X*' *presupposes* my knowledge of what I intend, rather than explaining it. But on the second use of 'will *X*', where I have no idea already of what '*X*-ing' might mean, 'will *X*' gives me no clue at all of what my intention might turn out to be. Boyle's comment on this is as follows:

> Part of what is odd about the idea that I might infer propositions about my present intentions from blank future propositions about myself is that it seems to get matters backwards. In certain instances, it seems, I believe that I will *X* precisely because I (knowingly) intend to *X* … The idea that the line of epistemic dependence runs in the other direction expresses a profoundly alienated picture of my knowledge of my own intentions—as if I must conclude to my own commitment to *X* from an unaccountable inkling about what I will in fact do. A subject who had to discover her intentions in this way would at best know of her own intentions; she would not know them through seeing herself in them. For her knowledge that she intended to *X* would not be grounded in her knowing commitment to *X*-ing. She would not know her intentions through seeing certain things as to-be-done. (2011, p. 234)

In this argument, by pointing out the alienated picture that the production-

process account of self-knowledge gives us, Boyle is also saying how lacking in commonsense it is.

Commonsense also prevails in the fact that reflectivism turns 'what we know' into an adverb. Instead of 'I know that I believe that *p*', we have 'I knowingly believe that *p*.' Thus, what used to be thought of as second-order knowledge of something *else* mental, hypostasising self-knowledge as a something (a mental object?) that has its own mental object is now only a first-order belief that is held in a particular way: that is, knowingly, and whose object is external and physical. This undercuts discussion about the relation between first- and second-order mental states.

Thirdly, and most importantly, given that this thesis is about estranged attitudes, reflectivism also gives us a way of understanding how it is that we can sincerely deny having a belief that we do in fact have. Reflectivism helps us to understand estranged attitudes, such as in self-deception and akrasia, together with the ordinary ways in which the availability of knowledge to our conscious awareness of it can be interfered with, such as by 'distraction, confusion, or temporary inhibition of memory' (p. 229), as well as, as I argue in Chapters 5 and 6, certain kinds of affect such as desire and fear. This is because reflectivism claims that *all* intentional attitudes, including estranged attitudes, in being held knowingly right from the start even when they are known *unreflectively*, are therefore *known*, albeit (when estrangement is present) tacitly only.

All of these are ordinary, standing attitudes. Every belief we form may become a standing attitude by being stored in memory, from whence it can normally be recalled when needed but from whence in some cases it may *not* be

accessible because of some factor such as one of those just listed.[10]

The key to grasping Boyle's position is to see that to be known tacitly is not necessarily to be known in a consciously aware way. This affects how we can understand Moore's paradox, for example (the paradox that arises from the fact that attributing a specific mistaken belief to oneself seems absurd; see Almeida 2001; Green and Williams 2007). Someone who avows i) 'Sea levels are rising' but refuses to avow ii) 'I believe that sea levels are rising' may know i) with conscious awareness but know ii) only tacitly. For example, it may be that she would feel disloyal to the political party she supports if she admitted to knowledge of belief ii). In all cases of estrangement, I am knowingly affected by an attitude that I hold, but I hold this knowledge only tacitly (Boyle 2011, p. 238). The estranged attitude itself can become active in me, but I am not active myself with respect to it while it remains estranged; I am passive, as when I suffer pain.

Self-deception and akrasia, on this view, are only a step away. When Barnaby sincerely but falsely avows, 'I am not in love with Clarice', he knows, unconsciously, that its opposite, 'I am in love with Clarice', is true. But he knows this second fact about himself only tacitly. Something is blocking this tacit knowledge from reaching his conscious awareness, and this blockage prevents him *in principle* from reflecting on it. Of course, if he discovers consciously that he is in love with Clarice, this knowledge cancels out its opposite. He can no longer believe consciously that he is not in love with Clarice. In most cases the 'something' that is blocking his knowledge will be a motivating emotion, such

[10] Boyle's reflectivist claim applies only to those attitudes that fit the Kantian category of pure apperception, signifying the 'I think' of active agency (similar to Aristotle's category of energeia, to be explained shortly) rather than that of empirical apperception or inner sense, where the agency involved is passive. The concepts of belief and of the other intentional attitudes arguably belong in this category, but a case would need to be made separately for each.

as, in Barnaby's case, fear of being rejected; and this emotion, expressible as a belief such as 'I am afraid she will reject me' is also tacit only.

Akrasia is also explicable on the reflectivist view. The chronic gambler can avow 'I do not intend to gamble tonight' while knowing, at the same time, that he does intend to gamble tonight. It's just that he knows this second fact only tacitly. He also knows, again only tacitly, why he intends to gamble. It is (say) *to win*, and thus perhaps to feel worthwhile for a time rather than worthless. But the moment he acknowledges this consciously, he confronts the logic of its irrationality and thus the consequences of losing yet again, and once he does *this*, he cannot gamble and so he cannot win. Therefore, he cannot acknowledge his real reason for gambling.

Boyle summarises his own position by saying:

> Thus it may be true that a person's believing *p* involves his knowing himself to believe *p*, and yet that a person can believe *p* without being conscious of it. This is not a paradox, so long as we are careful about the difference between being known and being accessible to conscious reflection. What is known is accessible to conscious reflection, other things being equal, but other things are not always equal. (p. 230)

Believing that *p* may involve knowing oneself to believe that *p* and yet not knowing oneself to believe that *p*. Why is this not a paradox? It is not a paradox because from the first-person position one may hold only tacitly that one believes that *p*. Indeed, one can sincerely avow, 'I do not believe that *p*', where this latter belief is explicitly held as true and can be consciously reflected on while at the same time, 'I believe that *p*' is held only tacitly.

It is important to see that people often operate in a tacitly deliberative way, when this way reflects their unconscious attitude rather than their explicitly conscious attitude. At such times, their non-verbal behaviour is a more reliable

sign of what they really think than their self-ascription is. However, this is not a sign of rational thinking, but rather of what the subject thinks is in her best interests. Desire and fear tend to motivate estrangement. Rational agency is interfered with, rather than enhanced, by the intrusion of estranged attitudes. Functioning well as agents requires either not being in the grip of an estranged attitude or being able explicitly to acknowledge that one has such an attitude but being able to control it so as to remain rational about that particular situation.

Conclusion

If reflectivism is viable, can we interpret Moran's account of self-knowledge in a reflectivist way to explain self-deception and akrasia? In Chapter 5 I argue that we can do so via the concept of estrangement. Repressed beliefs can remain active while the subject remains estranged from them. It is hard to see how this can be the case if they are simply empirical states that are deposited in our minds and then sit there like stones until they are reconsidered via memory. Let us set this out in more detail before we close this chapter.

According to reflectivism, we have tacit knowledge of all our beliefs: a person who believes that *p* 'is tacitly cognisant of being in this condition' (Boyle 2011a, p. 228). To have tacit knowledge of one's belief that *p* is to know that one believes that *p* without necessarily having ever become consciously aware of having this belief.

If reflectivism is viable, a self-deceived subject:

1. believes that *p*

and

2. believes that ~*p*.

There is no paradox involved in this as long as the subject is not consciously aware of one of these propositions. If she is not aware of 2, she can sincerely self-ascribe 'I believe that *p*'; and if she is not aware of 1, she can sincerely self-ascribe 'I believe that $\sim p$'. She just cannot self-ascribe beliefs 1 and 2 at the same time, from the first-person perspective, without paradox. This implies that we can sincerely deny having a belief that we do have. It explains how we can sincerely avow one of these beliefs and sincerely deny the other. So Barnaby can sincerely avow

1a. 'I am not in love with Clarice'

when, in fact, he sincerely believes (unconsciously)

2a. 'I am in love with Clarice'.

If Barnaby is estranged from his belief that he is in love with Clarice, he cannot, even in principle, self-ascribe that he is in love with her. He can, however, sincerely self-ascribe

1a 'I am not in love with Clarice'.

But 2a is also true of Barnaby: he is, in fact, in love with Clarice and, what is more, he knows that he is. The important thing, for the purposes of this thesis, is that he does not know it explicitly and consciously; he knows it only tacitly. This is because his knowledge that he is in love with Clarice is inaccessible to his conscious awareness. Reflectivism, in explaining how we can know what we believe, also explains how estrangement is possible. We are estranged from our belief when we know this belief only tacitly and cannot become consciously aware of it, even when we most need to.

Reflectivism thus provides the link between first-person authority and estrangement in self-knowledge, on Moran's account of it, that this thesis requires in order to argue that estrangement motivates many anomalous

actions. We can see that estrangement might well be produced by repression simply by asking *why* a subject would ever repress a belief. There are several commonsense reasons why Barnaby might repress his love for Clarice, such as that he fears rejection, a fear too painful for him consciously to reflect on. We can also see why this might motivate his anomalous self-ascription of

1a. 'I am not in love with Clarice.'

This self-ascription is anomalous because it does not make any sense when considered alongside the evidence of his devoted behaviour towards Clarice and the adoration in his eyes when he looks at her. But it makes sense to suppose that his need for emotional safety might well motivate the belief that it is safer not to be in love than to be in love. This example demonstrates how important it is for us to discuss the anomalies of self-knowledge; to be unable to express, even to yourself, your love for another person is an appalling predicament to be in.

In arguing for Boyle's reflectivism, I have tacitly supported Moran's position on first-person authority and estrangement: not only can we know our intentional attitudes immediately and with authority by making up our own minds about them in conformity with transparency, but we can also know our *anomalous* attitudes, those we have without being able consciously to become aware that we have them. We can know these latter by knowing them only tacitly; we cannot know them explicitly because we have repressed them, to use Freud's phrase, or have become alienated from them in some other way.

Chapter 5: Self-deception

Introduction

We have now reached the point where we can apply Moran's account of attitudinal self-knowledge to the anomalies of self-deception and akrasia, in order to show that his account explains them better than other accounts do. Other accounts tend to treat self-deception and akrasia as though they are practical irrationalities rather than failures of self-knowledge. In doing so, they ignore some obvious conceptual connections between self-knowledge and its anomalies. For example, self-deception and self-knowledge are conceptual opposites, and so are irrationality and rationality. Practical irrationality is often explained as a character deficiency in the subject, as though this has nothing to do with the person's knowledge of her own attitudes; and yet rationality, conceptually linked as it is with irrationality as its opposite, is central to many accounts of self-knowledge.

Moran's account gives us a new way of understanding self-deception and akrasia by conceiving them as failures of the first-person authority of self-knowledge. In his account, the first-person perspective on self-knowledge is irreducible to the third-person perspective—or, to put it a different way, Kantian active agency is irreducible to passive agency. This irreducibility enables us to distinguish between self-knowledge that is *both* epistemic *and* normative because the subject can endorse it as true or choice-worthy from the first-person perspective, and self-knowledge that is epistemic only, being no more than a third-person statement of psychological fact about what intentional attitude the subject has, regardless of whether it is true (if it is a belief) or choice-worthy (if it is a desire, an intention or an affective attitude). Once we have distinguished the two kinds of self-knowledge in this way, as

Moran's account does, we can explore what their differences imply for questions about self-deception and akrasia in a way that relates these to a lack of first-person authority.

In this chapter and the next, I discuss the anomalies: self-deception in this chapter and akrasia in Chapter 6. In this chapter on self-deception, in Section 1 I give a brief description of how to define it and how to understand the main ways by which philosophers have sought to explain it and why their suggestions are problematic. In Section 2 I question two accounts of it, those of Mele (2009, pp. 55–69) and Sanford (1988, pp. 157–169). Mele and Sanford argue that self-deception essentially involves motivated false belief, not necessarily involving any estranged attitude or any pair of explicitly incompatible beliefs, conscious or unconscious. I argue that neither Mele's nor Sanford's view is viable because in the end there is an estranged attitude involved that remains unmentioned. More specifically, I argue against Mele that in his examples of motivated false belief, his subjects' self-ascriptions do not conform to Moran's transparency condition. This gives us the evidence we need that such self-ascriptions contain an attitude that is about the subject rather than the object of the belief or other attitude she is self-ascribing. They are therefore biased, either consciously or unconsciously. I argue that in self-deception their bias must be unconscious; otherwise, they fall foul of the static paradox.

It needs explaining why Mele's sources of 'unmotivated' bias—that is, vividness of information and the confirmation bias—occur *as they do* and are used as they *are* used in the particular cases I discuss. I argue that an estranged attitude is motivating these sources to be used in this way.

Lay hypothesis testing fails to show why costly errors of one sort rather than of some other sort are minimalised, suggesting once again that an estranged attitude is responsible for this.

Sanford claims that self-deception is caused by mistaken desire structures. I argue against Sanford that we need to ask why a given subject has a particular mistaken desire structure in connection with some particular matter. If the subject is unaware that or why he has this mistaken desire structure, he is estranged from this knowledge. I conclude that in all of Sanford's examples, the subject has two contradictory attitudes, one conscious and one unconscious, and that these conflict, so producing the examples he discusses. The evidence for this conclusion is that in all of Sanford's examples, the subject cannot conform to transparency.

I therefore argue that neither Mele's nor Sanford's view is adequate because in none of their examples that I discuss does the writer sufficiently explore the subject's situation. In every case he overlooks considerations that indicate an estranged attitude and stops short of asking why when the answer would expose the inaccessibility of the reason in question. But this has to be teased out in each example. I must sometimes rely on claiming that it is more likely than unlikely that the subject has an estranged attitude. If she does, then, of course, she fails to conform to transparency. If I am wrong and she meets transparency, then on Moran's account, self-deception has not occurred. But a deflationist can insist that this is evidence for his own position, and with that I can only disagree. I present Moran's account as providing us with an explanation that if viable, is more reasonable. In Section 3 I briefly reinforce my own position.

Before continuing, I want to mention the thorny problem of defining self-deception. The traditional way of approaching self-deception models it on interpersonal deception, where you get someone to believe that p in order intentionally to deceive him. You believe that $\sim p$, but you do not want your victim to believe that $\sim p$. But this model faces the difficulty that you cannot do this to yourself, simply because you cannot get yourself to believe that p

without knowing that that is what you are doing, so that you are not really deceiving yourself at all. You already believe that ~*p*, and you cannot rationally believe both ~*p* and *p* at the same time. This is called the static paradox, and it does seem to many writers to be unsolvable. (Of course, if either of your beliefs that ~*p* or that *p* is unconscious while the other is conscious, that is another matter.) The intentionality involved in the traditional approach also presents difficulty. How can you intentionally get yourself to believe that *p* when you already believe that ~*p*? If you know what you are doing, then surely you are not self-deceived. This is called the dynamic paradox. Again, it seems to many philosophers to be unsolvable.

These paradoxes have led some philosophers to believe that self-deception is not possible. Others, however, have attempted to solve the problems of the paradoxes. Some have continued to argue that self-deception is intentional while others have denied this. One way of continuing to believe that self-deception is intentional is to suggest that within the self there is a division between a non-deceiving part and its deceiving part. The deceiving part might be a relatively autonomous sub-agency or a separate centre of agency or perhaps there is just a boundary between deceiving and non-deceiving parts.[1] In this way, they are suggesting that the non-deceiving part of the self might not be able to discover what is going on in its deceiving part.

Non-intentionalist writers, on the other hand, tend to explain self-deception as motivationally biased belief, so bypassing the problem of intentionality.[2] For example, they point out that you need not be certain that you believe that ~*p*; you may have no inkling at all of any problem. On this conception of self-deception, the self-deceived person need not entertain contradictory beliefs.

---

[1] Rorty, 1988, Pears, 1984 and Davidson 1985 have argued in one of these ways.

[2] Writers who have taken this approach include Barnes 1997, Johnston 1988 and Mele 2001.

But are you self-deceived in such cases? I will argue that you are not.

Since there are a number of approaches to this topic, there is no one definition of self-deception that all or most writers on this topic accept as correct. Thus there is no neutral definition for us to take as our starting point in this discussion. The commonsense definition of self-deception that I begin from in this chapter is as follows:

Self-deception necessarily involves two contradictory propositions, one of which is conscious while the other is inaccessible to the subject (the subject is estranged from explicit knowledge of it), together with an ostensible, rationalising attitude that hides the subject's real reason from herself and others.

This is a 'let us suppose' suggestion. I attempt to justify it as we proceed. Bear in mind in what follows, however, that Mele and Sanford both reject elements of the commonsense conception.

Section 1: Mele[3] and Sanford

Mele

Mele's (2009) first example is of Rex, who receives a rejection notice on a journal submission. Hoping that the rejection is unwarranted, Rex decides, on reading the referees' comments, that they have misunderstood two complex points he has made. A few days later, in a more impartial frame of mind, he rereads the comments and realises that the referees were right. The implication is that earlier, Rex is self-deceived: he thinks, falsely, that the referees have misinterpreted his complex points. He thinks the rejection is not justified.

[3] Mele, 2009.

Is this sufficient for us to agree with Mele that Rex is self-deceived at this earlier time? The claim is that he is self-deceived because his desire to have written an article that is fully acceptable has biased his thinking towards believing that this article really is of publishable standard. But to convince himself of this, he has to invent a rationalisation: the referees have misunderstood two complex points he has made. This rationalisation is false. But as long as he can believe it, he can remain self-deceived.

I suggest that the article's rejection has temporarily knocked Rex's thinking off balance. Mele continues his description of this example by saying that 'a few days later, when Rex rereads his paper … in a more impartial frame of mind', he realises the referees were right. Mele claims that examples such as this one are 'garden-variety instances of self-deception' (p. 57).

Moran restricts his account to settled attitudes. There is nothing settled about Rex's very temporary belief that his referees were wrong. He doesn't even reread their comments until three days later, at which time he realises he was wrong. All he needs to do is to be able to reread their comments 'in a more impartial frame of mind' (p. 56). If cases like that of Rex are cases of self-deception, we are all self-deceived a great deal of the time, whenever something upsets our mental equilibrium for a while. But self-deception requires, on most definitions of it, that the subject has a motive for believing her ostensible belief. Rex's motive is to avoid the anxiety he feels at reading the referees reports, the anxiety about being good enough to continue in his university career. This is not a short term motive.

Audi, for instance, claims that 'the dynamics of self-deception requires, at least normally, a gradual onset' (1988, p. 96). Audi believes that self-deception is never identical with an act but rather with patterns of behaviour. This fits extreme cases such as those healable only by such measures as psychotherapy, quite well with most medium term self-deceptive acts and not so well with very

short acts of supposed self-deception such as that of Rex's with his referees. Of course, one can have a long-term ongoing fear that (say) one's partner will die soon, such that when one hears bad news about this partner one immediately believes the worst. But this is because a long held motive suddenly provides a reason for the subject to believe what he has long feared. In this case, it is likely that this belief will take at least several weeks, shall we say, to disperse.

Let us assume first that Rex does not have two contradictory beliefs. This is not a criticism of Mele because the thrust of Mele's argument is that self-deception does not require two contradictory beliefs. But he recovers his equilibrium too fast for someone who is self-deceived. It seems to me that if he suffered from self-deception about this matter, Rex would continue to believe for some time that the referees were wrong.

We can understand him as being self-deceived if we take the view that in the short time during which his thinking was off balance, he was literally unable to consider the possibility that he might be being biased by his desire to have the article published. This would be to say that during that time, his desire was inaccessible to him and that would be to say that self-deception can occur with almost lightning speed in certain situations. As I have said, I would prefer to limit self-deception to cases where the self-deceptive belief is settled. If Rex had believed for some months that his article was of publishable standard then his belief that the referees were wrong could occur with lightning speed as he read their reports, and I would be more likely to think that since he immediately said, 'I believe the referees misinterpreted my complex points' (this being his ostensible rationalisation), he was straight away self-deceived. However, in that case, this self-deception would take more than three days, I would think, to subside. As to what length of time a self-deceptive belief must last to count as settled, I think that would depend on the example.

Might we also find two contradictory beliefs in this example?

I suggest they are these:

A: My article is of publishable standard.

B: My article is not of publishable standard.

Rex believes A consciously and firmly. But unconsciously, let us suppose he believes B.

Why does he believe B? Let us say he believes B because he believes he is not good enough to get something published yet. This belief, unconscious though it is, causes him much angst. This angst causes him to feel desperate about his situation—he desperately *wants* to believe that he has reached publication standard. (Perhaps he fears irrationally that otherwise, because of his immigrant background, he will miss out on university employment.) The desperation motivates his bias towards believing the opposite of B. This is A: that his article is of publishable standard. Let us assume that the referees have not misinterpreted his two complex points. In that case, Rex may be self-deceived if he is not consciously aware that his rationalising belief A is biased by his desire that his article be of publishable standard.

This raises the static paradox of self-deception. Rex cannot consciously believe both A and B at the same time. It is only while he remains estranged from B that he can continue to believe that A is true. As soon as he acknowledges the possibility that his thinking has been biased by his own desire in favour of believing this, he cannot continue to believe that A is true. He can continue to *hope* that it might be, but that hope has now become nothing more than wishful thinking.

One reply to what I have just argued is to say that once Rex acknowledges that his desire might have been biasing his thinking, he can continue to think that

same way because he can do so *irrationally in practice*. After all, is not self-deception an example of a practical irrationality?

But you cannot continue to think irrationally in this particular way once you suspect that this thinking has been biased by your own desire. In this example, the evidence that Rex does not continue to think irrationally is that he does *not* continue to believe that the referees have misinterpreted his complex points. Being able consciously to acknowledge the suspicion has made his frame of mind more impartial—he can now be more realistic and objective. He re-reads his point more carefully and can now see what he could not see earlier: the referees were right. Earlier he could not see this, because *unknown to him at that time*, his desire was biasing his thinking. It follows that he was, at that time, *estranged from* the thought that his desire to have achieved publishable standard in his article was biasing his thinking about his complex point.

Rex might consciously fear, at the earlier time, that his desire *might* be biasing his thinking. But if you had asked him about this at that time, he would have said that on the whole he believed it was not. The evidence for this is that, early on, he sincerely believed that the referees had misinterpreted his complex points.

My conclusion at this point, based on this one example, is that what you take to be an example of self-deception may depend partly on its degree of irrationality together with the length of time you think the subject must remain irrational in order for your current example to qualify as one of self-deception. But let us continue.

I think that the static paradox rears its head more clearly in Mele's second example. This is of Sid, who wants Roz to love him. Sid misinterprets Roz's declining his invitations and reminding him that she has a steady boyfriend as evidence that she is playing hard to get in order to make Sid prove that his

love for her is as strong as hers is for him. Certainly he is self-deceived about her. But Mele's explanation of Sid's behaviour as simply wanting Roz to love him is too facile. Sid's misinterpretations continue for some time, possibly even amounting to harassment, indicating a deeper motivation. I suggest he has an inaccessible fixation on Roz that he would deny if asked. Again, we might say that it is his irrational attitude that explains his self-deception. But then I would want to ask why he has this irrational attitude. Unless we postulate a reason for this, we cannot fully understand his ongoing persistence. I suggest that his behaviour shows a desperation from which he is estranged. He does not behave as he would if it were *just* that he wants Roz to love him. He must long for her love more desperately than he is aware of. Again, the evidence for this is that he continues to invent highly irrational explanations, such as that she is playing hard to get, in the face of her presenting him with facts that are inconsistent with this. He must in fact believe that she is lying when she says she has a steady boyfriend. This is *very* irrational! He simply does not take clear evidence into account over a solid stretch of time. The reason is, I suggest, that if he lets himself consciously *feel* his desperation he will realise its irrationality for himself. The static paradox then comes into operation: he must acknowledge to himself that Roz is not lying: she actually does have a steady boyfriend. Unless he is too deeply fixated on her to care, in which case he suffers from delusion, he must reconsider his situation.

This example supports Mele's (and my own) position that self-deception is not intentional, but it seems to me that there is plenty of evidence in it that Sid does have an estranged attitude towards Ros. It seems to me that he believes:

A: Ros loves me; she is just playing hard to get.

and

B: Ros does not love me.

A is fully conscious. B is unconscious. Sid holds both beliefs at the same time.

Mele's next example (in 1909) is of Beth, whose father has recently died and who wants her father to have loved her more than he loved her brothers. She misinterprets the photographs she looks at so that they seem to show that her father loved her more, when really they show her father's true feelings, which were that he loved his sons more. Mele indicates two sources of unmotivated bias in this example: vividness of information and the confirmation bias.

But here, again, we find the static paradox. Is Beth consciously aware of why she finds these photos pleasant? If she is consciously aware that it is because she wants to believe they show that she was her father's favourite child, then she is not self-deceived. If she does not know why she finds these photos pleasant and the family photo albums unpleasant, then she is self-deceived into believing that the photographs are evidence that she was her father's favourite child. I suggest she has the correct but inaccessible fear that her father did not love her as much as he loved her brothers but that this fear is inaccessible to her because it is too painful to acknowledge. Otherwise, why would she so grossly misinterpret photographs that are clear evidence that he loved her brothers more? If this is right, she cannot sincerely avow both of the following at the same time:

A: These photos show that my father loved me more than he loved my brothers.

B: My father loved my brothers more than he loved me.

Once she consciously acknowledges B, she cannot continue to believe A.

I suggest that vividness of information and the confirmation bias are ways of producing and maintaining bias in Beth's case. Mele lists negative and positive misinterpretation, selective focusing and attending to evidence, and selective

evidence-gathering, as ways by which we can become self-deceived *without* having any relevant estranged attitude (p. 56). But why would we approach the evidence for some event in a motivationally biased way unless we had some reason for doing so? To avoid explaining *why* a self-deceived subject has a certain motive, Mele argues that desire can be explained in terms of *unmotivated* bias, although he also accepts that biases may be motived. Desires, Mele argues, can enhance the vividness or salience of data, influence which hypotheses one thinks of and selects, and thus cause confirmation bias. But desires for what? And why do we desire this rather than that? Mele supports his argument with a short discussion of lay hypothesis testing, which claims that such biasing is driven by a concern to minimise costly errors rather than being driven by motive. However, what one conceives as a costly error to be avoided will depend on what one desires to prove or disprove. Mele quotes Friedrich, one of the formulators of the lay hypothesis testing theory:

> A prime candidate for primary error of concern is believing as true something that leads [one] to mistakenly criticise [oneself] or lower [one's] self-esteem. Such costs are generally highly salient and are paid for immediately in terms of psychological discomfort. When there are few costs associated with errors of self-deception (incorrectly preserving or enhancing one's self-image), mistakenly revising one's self-image downward or failing to boost it appropriately should be the focal error. (cited in Mele 2009, p. 59).

Mele tends to agree with this, pointing out that, for example,

> a strong desire to maintain one's relationship with one's spouse plays a role in rendering the potential error of falsely believing one's spouse to be innocent of infidelity a 'costly' error … and more costly than the error of falsely believing one's spouse to be guilty. After all, the former error may reduce the probability that one takes steps to protect the relationship against an intruder. (2009, p. 59)

This strong desire to maintain one's relationship might of course be conscious. But if it is conscious, the costly errors that Mele discusses may cause the subject

to become mistaken about his spouse, but not self-deceived. The evidence that he is not self-deceived is that he does not give an ostensible, rationalising reason for his suspicions: he gives his real reason, that he is worried that his spouse is having an affair. You cannot give your real reason when you are self-deceived because you are estranged from it. On the other hand, is this subject consciously aware of his need to preserve or enhance his self-image in cases where this is his real reason for believing his spouse to be innocent of infidelity? Surely not. Surely he is focused on the evidence and on his very conscious anxiety, instead of asking himself right now whether he is really only concerned with maintaining his own self-image. But maintaining his self-image, if this desire is *un*conscious, *might* well be his real reason for worrying about whether his spouse is having an affair. If his real reason is his distress at the thought of losing her, this reason is likely to be conscious and thus he is not self-deceived. Even so, if his distress is excessive, he might become unduly suspicious or even paranoid about her current absences from home. (Perhaps she is really taking French lessons, so that she can accompany him to France more confidently next time.)

The point is that if there is an unconscious attitude in this example (such as the subject's need to maintain his self-image or his estrangement from his belief that his real reason for insisting that his spouse is not being unfaithful is his strong need to maintain his relationship with her), it is not avoided by associating it with 'a costly error'. In the unlikely case that he is consciously aware of his need to maintain his self-image, he is not self-deceived in his suspicions if it is this need that drives them. If he does not find it so unacceptable that he has to repress it, then he can consciously accept it. In this case, he is mistaken about his spouse's intentions but not self-deceived about it. If this is right, the static paradox has reared its head again. For this to be self-deception, his real reason must be inaccessible to him and his rationalising ostensible reason, contradicting it, must be conscious.

Here is one more example from Mele's article. Art is angry with Bob for a recent slight. His anger leads him to view Bob's behaviour as more hostile than it is (p. 60). But I suggest that even if Art is aware of his anger, as the example implies, he is self-deceived in his misreading of the extent of Bob's hostility because he (Art) does not believe that it is his anger that is motivating his belief when it is. In this case, the degree and force of his anger are inaccessible to Art. He believes that Bob is very hostile, when in fact Bob is much less hostile. Mele argues that Art's anger 'might prime the confirmation bias by suggesting an emotion-congruent hypothesis about Bob's current behaviour—for example, that Bob is behaving badly again—and it may increase the salience of data that seem to support that hypothesis' (2009, p. 60). This might be true, but it does not change the fact that Art's misreading of the extent of Bob's hostility towards him is motivated by an anger that he does not accept as being its cause. He is therefore estranged from the damage that his anger is doing to his friendship with Bob. Again, the static paradox is present. Art cannot consciously believe both of the following at the same time:

A: Bob acted with a great deal of hostility.

B: Bob acted with mild hostility.

Once he discovers how much his anger with Bob is influencing his opinion about how hostile Bob is, he must re-evaluate his assessment of Bob's behaviour. He can no longer believe A.

So far I have suggested that Mele's position fails to explain self-deception adequately. In the cases just discussed, it is more likely than unlikely that the subject does have an estranged attitude because if she does not, the static paradox is present. Or, the subject is not self-deceived because the example is too trivial.

We can reinforce this conclusion by asking whether, in each of the above

examples, the subject's self-ascription conforms to transparency. It cannot do so when the subject has an estranged attitude. This estranged attitude, I conclude, is a necessary condition for self-deception. It is not a sufficient condition: an ostensible, rationalising belief is also necessary.

I want to repeat, at this point, my understanding of what failure of conforming to transparency implies for the subject. Does it imply that from a third-person point of view, the subject has only third-person authority over her self-ascription? We can best decide this by considering the following quotation from Moran:

> A fairly modest version of the idea of 'first-person authority' will understand it not as entailing either infallibility or perfect access, but as a feature of discourse, as the authority a speaker is ordinarily granted to declare his thought and feeling, and have that declaration count (normally decisively) as telling us what the person's attitude is. (Moran 2001, p. 121)

How does this help us to understand what happens to first-person authority when a subject 'avows' her rationalisation of her self-deceptive belief? If first-person authority is only 'a feature of discourse', and not to be taken as infallible or as her describing of perfect access to her mind, then we can say that in this particular case, the authority need not be taken as gospel. It must be at least queried. It does not usually tell us, for example, that the subject has a relevant estranged attitude. When this is the case, it does not mean that the subject has not spoken from the first-person perspective she has. It is just that her estranged attitude has biased her belief. But first-person authority has still done its job: it has told us what she believes, even when that belief is irrational, even when, from the third-person perspective, it is false. Objectively speaking, we can say that her *attempt* at avowal gives us no more than third-person, theoretical knowledge of her attitude. We might say that objective, theoretical knowledge of whether what she believes is true or false is not the job of first-

person authority, its job is to tell us what the subject (rightly or wrongly) believes. It is often important for others to know this. Indeed, it is essential to the whole business of rational discourse for us to be able sometimes to access another's mind in this first-person way. However, it might be that when the ordinary assumption of first-person authority is questionable, and we find that the subject cannot give a satisfactory reason for her avowal, we assign it only the very ordinary authority of an attribution from the third-person perspective.

Sanford

Further to pursue my claim that deflationary approaches are less reasonable than my own, I turn now to Sanford, who argues that self-deception is caused by mistaken desire structures. My first example from Sanford is that of the baseball enthusiast. In this example we need to keep in mind Sanford's claim that self-deception requires only a mistaken desire structure. He sets out what it does not require as follows:

> … there need be no reason to suppose that the self-deceived person 'really knows' things to be other than he believes them to be. Self-deception does not in general require belief in the face of evidence or a conflict between what one in some sense knows and what one believes. (1988, p. 162)

Sanford denies here that self-deception requires a belief at odds with the evidence or conflict between what one knows and what one believes. All that is required is a mistake in the subject's desire structures. Such mistakes, Sanford points out, are common. But I argue now that Sanford's baseball enthusiast example requires, to make it viable, contradictory beliefs (*p* and *~p*), one conscious and the other unconscious, and conflict between them, such that any relevant 'avowal' he might make will not be fully psychologically free because of an estranged attitude which is preventing the subject from conforming to transparency. Here is the example.

The baseball enthusiast

> I think my daughters would really enjoy going to tonight's baseball game. This is my reason, I think, for wanting to take them to the game. But the real order of dependence is the reverse of this. While I am not so self-ignorant that I fail to realise that I would like to go myself, I do not quite realise that I am unlikely to indulge this desire without some additional reason for doing so. My belief that they would enjoy going serves to justify my acting on my desire to take them. They do not manifestly hate going to baseball games, so my belief is not absurd. It may even be true. The point is that I have it because it rationalises a desire of mine, while I mistakenly think that I have the desire because I have the belief. There may well be lots about my desire of which I am ignorant, but there need be nothing that I really know but am trying to cover up. I do not really know, in particular, that I think they would enjoy going to the game because I would like to have an acceptable reason for taking them. (Sanford 1988, pp. 162–163)

Sanford uses this example to demonstrate his claim that 'there need be no reason to suppose that the self-deceived person "really knows" things to be other than he consciously believes them to be' (p. 162). He claims that self-deception does not in general require belief in the face of evidence or a conflict between what one in some sense knows and what one believes (p. 162). Let us see if this is the case here.

Consciously, our baseball enthusiast can 'avow', 'I am taking the girls to the baseball game because they would enjoy it.' But he also knows, consciously or unconsciously, that he is going to the baseball game mainly because *he* wants to go. So the baseball enthusiast has a mistaken desire structure. He thinks that he wants to go to the game because he believes his daughters would enjoy the game; but really, he takes his daughters to the game because he wants to go himself and needs an additional reason for going. So he has mistaken *the ordering* of his desire/ belief structures. He thinks his belief (that his daughters will enjoy the game) causes his desire (to go himself) when actually his desire

(to go himself) causes his belief (that his daughters will enjoy the game). Sanford implies that both the belief and the desire can be conscious: the baseball enthusiast just gets them the wrong way round. He deceives himself about which causes which, but both are conscious and there seems not to be any conflict between them.

But Sanford stipulates in this example that there is nothing that he knows only tacitly, *unconsciously*. ('I do *not really know*, in particular, that I think they would enjoy going to the game because I would like to have an acceptable reason for taking them' [my emphasis]). He does not, he stipulates, have that thought unconsciously and thus 'tacitly' knows it. However, he does really know, tacitly and unconsciously, that very thing. What he knows unconsciously is that he wants to have an acceptable reason for taking his daughters *so that he can go himself as well*. But why is he 'unlikely to indulge this desire without an additional reason for doing so? Why does he not want to go by himself, without having an 'acceptable' reason? This is what we are not told. Might taking his daughters be 'an acceptable reason' because (say) it would not antagonise his wife or avoid his having to mark his exam papers that night? There could be a whole host of reasons. If it is either of these, then his duty to his wife (keeping her company?) or to his job is his real reason for wanting to take his daughters with him to the game. So he does have an unconscious reason (to not antagonise his wife or to avoid having to mark his exam papers) for inviting his daughters to the baseball game.

It follows that there are two contradictory reasons in this example, one conscious and one unconscious, and that they are in conflict. His conscious reason is that he is going to the game to please his daughters as well as himself. His real but inaccessible reason is that he is going to please himself and must take his daughters with him so as not to upset his wife or neglect his job. These are in conflict. This means that if he 'avows' his conscious reason (his ostensible

reason), the 'avowal' will not conform to transparency. If he 'avows', 'I want to go to the game because my daughters will enjoy it' this is false, and is opaque to his real reason. His unconscious reason is mildly manipulative of his daughters whereas his conscious reason seems, on the surface of it, to be a kindly, fatherly gesture. He seems to have no compunction about involving his daughters in this very mild and non-damaging way, so it seems his conscious reason has won the conflict.

The example therefore fails to show that 'one can be self-deceived without hiding from some truth that one really knows'. The baseball enthusiast successfully hides his tacit knowledge of his real reason for wanting to take his daughters with him to the game and to do so he invents a false, rationalising reason. If cases as mild as this can be cases of self-deception, then the baseball enthusiast is self-deceived.

Because he is self-deceived, he cannot believe both of the following reasons consciously at the same time:

A: I do want to go to the game myself, but that is not why I am taking my daughters there—I am taking them there because I know they will enjoy it.

B: My real reason for taking my daughters to the game is that I want to go myself and need an acceptable reason for doing so.

The static paradox now cuts in. He cannot continue to believe A once he has consciously admitted to B. Once he consciously acknowledges B, he knows that he has been deceiving himself about his real reason. He is no longer self-deceived.

My other example from Sanford concerns a stamp collection. Sanford argues here that

> being self-deceived consists in one's misapprehending the structure of one's

> attitudes, in one's taking the having of one attitude to explain the having of another when the true explanation is something else. Such misapprehension does not require inconsistent beliefs or a belief in conflict with what one really knows. (1988, p. 169)

I argue that he cannot be right about this.

The stamp collection

The gambling debts of Jones's son, Sonny, had become very worrisome to everyone in the family, including Sonny. One day both Sonny and his car were gone. So was Jones's valuable stamp collection. There has been no word about the stamp collection, the car or Sonny. Jones says he does not believe that Sonny took the stamp collection.

Jones is reluctant to admit Sonny is a thief and in particular he is reluctant to admit that Sonny would steal *from him*. He knows all the evidence and where it points. He admits the evidence is that Sonny stole the stamp collection. He admits that it is a reasonable belief. He says he does not accept the conclusion himself. He does not believe that Sonny stole the stamps. He admits that this is an unreasonable belief but he says that it is not in his power to give it up.

Next, Jones is told by the police that Sonny has been arrested in a nearby city for attempting to sell the stamp collection to a dealer. He now says that he deceived himself about Sonny's innocence. He admits that the others were right, they really knew all along that Sonny had stolen the stamp collection but he says he did not know this himself.

When he finally comes to the sad realisation that Sonny did steal the stamp collection, he admits that he deceived himself.

Sanford contends that being self-deceived consists in one's misapprehending the structure of one's attitudes, in one's taking the having of one attitude to

explain the having of another when the true explanation is something else. Such misapprehension does not require inconsistent beliefs or a belief in conflict with what one really knows.

But Jones does have an estranged belief. He admits that his belief that Sonny did not steal the stamp collection is unreasonable and he does not have the power to give it up, but says that he does not know why he does not have the power to give this belief up.[4] If he does not know why he lacks this power then he is estranged from his real reason for lacking it, particularly after having acknowledged that it is unreasonable.

His estranged attitude is, I suggest, that although he knows theoretically that Sonny stole the stamps *from him,* he cannot admit this fact to himself because it is too painful for him to believe this about the son that he loves. In fact, he is still estranged from its painfulness in the final version of the story and is thus still self-deceived. He says that he had 'held a belief in the teeth of evidence whose strength he estimated correctly'. Well yes, he did, but why did he hold this belief in such circumstances? All he can say about this is that he did not have the power to give the belief up. We need to take a further step and ask why he did not have this power. I suggest that his admission of not having the power to give up his belief that Sonny did not steal his stamp collection was his ostensible rationalising reason, given to protect himself, unconsciously, from the pain that would otherwise have accompanied his acknowledgment that Sonny stole the stamp collection *from him*. He knew by now and had known *theoretically* for some time that Sonny did steal the collection, but he could not let himself feel the painfulness of this *for himself*. He therefore could

[4] Note that the reason for Jones's lack of power to give up his unreasonable belief is an estranged attitude. This supports the position on willpower that I take on serious cases of akrasia, in Chapter 6.

not endorse this belief and thus his early self-ascription, 'I believe that Sonny stole the stamp collection', could not conform to transparency because at the level of conscious awareness he could not bring himself to believe that Sonny really had stolen the collection *from him*.

If this is right, then Sanford's definition of self-deception as misapprehending the structure of one's attitudes is not viable. Self-deception must also include an unconscious attitude, one that in this case is too painful for Jones to admit to. Jones has an estranged attitude: that is, the unbearable painfulness of his knowledge that Sonny stole the stamp collection *from him*. This has cut him to the heart, and has had to be avoided. The same consequences follow: Jones has two contradictory attitudes, one conscious and the other unconscious, and these conflict. Jones' 'avowal' cannot conform to transparency. The static paradox is present again. Jones is self-deceived as long as he continues to believe both of the following:

A. It is reasonable to believe that Sonny stole my stamp collection.

B. I do not believe that Sonny stole my stamp collection.

His ostensible, rationalising belief, produced to protect him, is that he does not have to power to believe that Sonny stole the collection.

Once he acknowledges the painful truth he can no longer believe either B or his former rationalisation. But note that he has to *feel* the pain, not just acknowledge the reasonableness of what the pain is about. Moran's account of self-knowledge is practical as well as epistemic, involving a specific mode of awareness. He says:

> The problem of self-knowledge is not set by the fact that first-person reports are especially good or reliable but primarily by the fact that they involve a distinctive mode of awareness and that self-consciousness has specific consequences for the object of consciousness … a conscious belief enters into different relations with

> the rest of one's mental economy and thereby alters its character. We speak of the 'consciousness' in 'conscious belief' as something that informs and qualifies the belief in question, and not just as specifying a theoretical relation in which I stand to this mental state. (2001, pp. 28; 30 to 31)

Once Jones consciously accepts the painfulness of Sonny's behaviour towards him, he understands both Sonny and himself a bit differently. He is more realistic, more practical about their father/son relationship. This difference will be played out in various ways in their future interactions.

The stamp collection example serves as a counterexample to Sanford's central contention. Misapprehending the structure of one's attitudes must have a reason. Unless it is a 'brute' reason, such as fatigue, it normally does require inconsistent beliefs, one of which is inaccessible. Only when the inaccessible belief has been rendered accessible can the subject conform to transparency.

I have now argued that in both of Sanford's examples, the subject has two contradictory attitudes, one conscious and the other unconscious, and that these conflict, producing the examples he discusses. Therefore in neither example can the subject conform to transparency; in both examples the subject has an estranged attitude.

My argument so far introduces us to the fact that empirical accounts of self-knowledge cannot distinguish between the first-person, normative, avowable perspective and the third-person, empirical, attributive perspective. This inability, I argue, explains why deflationary accounts, whose aim is to reduce the first-person to the third-person perspective, cannot fully explain either self-deception or akrasia. Only an account from the irreducible, first-person perspective can allow that p and $\sim p$ can both be known at the same time without incoherence as long as one of these (p or $\sim p$) is known only unconsciously at that time. Moreover, these opposites can conflict with each

other in a way that can always be resolved in principle although only sometimes in practice. Deflationary accounts thus cannot provide the commonsense explanation of self-deception that is available to an account based on irreducibility.

I have argued so far that in the examples I have used, an estranged attitude, of whatever degree of severity, is essential. But this needs further discussion. We also need to discuss whether self-deception can be intentional. In Section 3 I discuss Audi's and McLaughlin's accounts, asking whether and if so how self-deception can be intentional and whether they believe that an estranged attitude is essential. Audi's and McLaughlin's accounts are both largely compatible with my own. In Section 4, I reinforce my own view on self-deception.

Section 2: Audi and McLaughlin

I begin by discussing Audi's (1988) necessary condition for self-deception, since I largely agree with it, together with one of his examples, followed by McLaughlin's (1988). I do this to provide exegesis, to compare these others' approaches with Moran's (from which they are not greatly dissimilar), and to place Moran's account in the field.

Audi

First, here are my necessary conditions for self-deception:

Self-deception necessarily involves two contradictory propositions, one of which is conscious while the subject is estranged from the other (it is inaccessible to her explicit knowledge at that time), together with an ostensible rationalising attitude, given to hide the subject's real reason from herself and others.

Audi's understanding of self-deception is reasonably similar to my own. He says:

> A person, S, is in a state of self-deception with respect to a proposition *p*, if and only if:
>
> 1. S unconsciously knows that *~p*;
>
> 2. S sincerely avows or is disposed sincerely to avow that *p*; and
>
> 3. S has at least one want that explains, in part, both why S's belief that *~p* is unconscious and why S is disposed to avow that *p*, even when presented with what he sees is evidence against *p* (1988, p. 94).

I agree with this, as long as we equate 'unconsciously' with 'inaccessibly' in point 1 and 'unconscious' with 'inaccessible' in point 3. Audi's comment allows us to do this most of the time. He comments:

> Here, unconscious belief is understood in a nontechnical and quite unmysterious sense. It is simply belief which S cannot, without special self-scrutiny or outside help, come to know or believe he has; it is not buried in a realm that only extreme measures, such as psychotherapy, can reach.[5]

However, I disagree with the last part of his comment. My own view includes, but is not restricted to, situations such as those that only measures such as psychotherapy can reach. Nor does Audi link his account to apparent failures of first-person authority as Moran's account does. He cannot therefore use the evidence of transparency in deciding whether the subject has successfully expressed her belief with first-person authority.

Audi argues that in self-deception there is a balance between the forces that dispose the subject to maintain her self-deception and the forces that dispose

---

[5] Audi footnotes this last statement by referring the reader to his 'The Concept of Believing', *The Personalist* 57 (1972) pp 365 - 377.

her to see the truth plainly, so becoming no longer self-deceived. Where this balance is lost and the subject slides into a state where she cannot know the truth at all, not even unconsciously, she has become deluded. I agree with this also. If Audi is right, self-deception manifests a degree of rationality, although not enough for the subject's 'avowal' to conform to transparency. Where the subject does really know the truth, albeit only tacitly, and thus does not wholeheartedly believe her rationalisation, 'the evidence, has, in an important way, prevailed' (Audi 1988, pp. 109–110). I mention delusion here because Audi argues that in delusion, the subject cannot know the truth at all, not even unconsciously. This gives us a point of contrast between delusion and self-deception (in which the subject unconsciously does know the truth) that suggests that self-deception is reasons-based and so can be conceptually connected with self-knowledge as the opposite of first-person authority, as Moran claims.

Let us consider one of Audi's examples, that of adolescent Jan. Audi sets it out thus:

> Suppose that Jan is an adolescent girl who has had an unhappy childhood, is subject to depressive moods and craves attention. She might 'attempt' suicide by taking an overdose of aspirin, say, six tablets. The result might be that her parents show alarm and begin to pay more attention to her. She might inform her friends about the incident, too, and perhaps tell them, when she feels low, that she is again contemplating suicide.

Audi presents the example of Jan as a case where it is very unclear whether she might commit suicide or not, because there is evidence in her on-going behaviour that points both ways. He comments that she is in a state of self-deception with respect to the proposition that she is seriously committing suicide. I take him to mean by this that she is in fact not serious about committing suicide but is self-deceived in thinking that she is serious about it.

He then agrees that her motive for committing suicide would be completely understandable (because of her unhappy childhood) and that it is also very understandable that she should believe that appearing suicidal is a way to get her parents' attention—such thoughts might be conscious or unconscious or sometimes the one and sometimes the other.

Let us consider this example further. Suppose Jan avows, 'I believe I might commit suicide this weekend'. We can find out whether she is self-deceived about this firstly by asking whether this avowal conforms to transparency. How do we do that? First we need to ask what the evidence is for her belief. We know what this evidence is: she has already tried and failed to commit suicide, and she has been exhibiting depression for some time. If her belief is true and genuine, and she is not using it as a way of getting her parents' attention or for any other reason that is about *her* rather than about the problem she is talking about, then her avowal is successful. It has first-person authority because it is a genuine belief based on reasons that are solely about its object (herself committing suicide) and is at least minimally rational.

If, on the other hand, she does not actually believe that she might commit suicide that weekend, then we need to ask why she 'avows' that she might. First, is her belief that she will not commit suicide that weekend conscious or unconscious? If it is conscious then she is not self-deceived. So let us suppose it is unconscious. In this case, she is estranged from it. So why does she 'avow' that she might commit suicide that weekend? Since she does not consciously believe this 'avowal', she avows it for motives that are inaccessible to her. The most likely motive would be that she wants her parents to prevent her from committing suicide. That would be acceptable as a real cry for help, albeit inaccessible. Or, her main aim, either instead or as well, might be her need to hurt her parents in revenge for not giving her the love and protection she needed as a child. This last aim also has to be unconscious, because otherwise

she is not self-deceived at all—she is simply lying. It is also, I suggest, her main reason for self-ascribing a belief that at the unconscious level she knows she does not have, even though at the conscious level she does believe that she might commit suicide that weekend.

To qualify as self-deception, Jan also needs to satisfy my second criterion for self-deception, that of giving a rationalisation. She might say, 'I cannot go on living any more'. This is meant to explain her avowal, and it is, of course, false in the sense that unconsciously, she knows that she can continue as before. But it would upset her parents even more, playing into her need to take revenge on them.

It is hard for us to accept that we may have bad desires unconsciously in our minds. We want to help and support Jan, not criticise her by suggesting that she may have such desires. In my view, this is one reason why estranged attitudes are so hard to discover. We think that if we have desires to hurt or destroy we are bad people. I think, on the other hand, they imply that we are ordinary people who are often unaware of feelings that we have been brought up to think of as unacceptable. I know several people quite well who never see anything bad in anyone. Anyone who suggests there might be some such desire lurking at the back of someone's mind is often condemned as being too judgmental. One reason why it is important to reclaim our estranged attitudes is that we can then begin to understand ourselves and others better, becoming hopefully less judgmental in the process.

Audi's account of self-deception works well for me on the whole. Three differences between us are that Audi does not apply the transparency condition, that he believes that self-deception is always a state rather than an act, and that he does not include cases requiring psychotherapy in his definition of unconscious attitudes relevant to self-deception.

Since McLaughlin partly supports Audi's position, this brings us to McLaughlin.

McLaughlin

McLaughlin (1988) asks how self-deception is possible. He agrees to some extent with Moran's link between estrangement and self-deception by giving us his own way of differentiating between accessible and inaccessible beliefs and by arguing that contradictory beliefs do cause conflict in the subject. I will largely use McLaughlin's own words.

Let us say that:

> *x*'s state *B* is an accessible belief that *p* at *t* if and only if (i) *B* is a belief that *p* at *t*, (ii) *x* can think of *p* at *t*, and (iii) were *x* to think of *p* at *t*, *x* would do so by means of *B*'s being manifested by *x*'s thinking that *p* at *t* … And let us say that *x*'s state *B* is a *consciously inaccessible* belief that *p* at *t* if and only if (a) *B* is a belief that *p* at *t* and (b) at *t*, *B* is not an accessible belief that *p*. It follows that if *B* is an inaccessible belief that *p* at *t*, then either *x* cannot think of *p* at *t* or else it is *not* the case that were *x* to think of *p* at *t*, *x* would do so by means of *B*'s being manifested by *x*'s thinking that *p* at *t* (that is not how *x* would think of *p*). In the second alternative, condition (i) obtains but (iii) does not. (pp 4 to 50)

Here McLaughlin claims that a subject can have an inaccessible belief at a particular time, and that the inaccessibility of this belief consists in the fact that the subject *cannot* think of it at that time. He also says:

> Someone who is self-deceived in believing that *p* may well think that *p*, sincerely say that *p*, and even defend the belief that *p*. The question naturally arises as to whether, when the belief that *p* is expressed, that would eradicate the belief that not-*p*. I maintain that it *need not* … What it can do instead is render … the belief that not-*p* inaccessible (pp. 50–51).

Here he claims that contradictory beliefs can be maintained at the same time in cases of self-deception as long as one of these beliefs is inaccessible to the

subject's awareness of it at that time. This is consistent with saying that beliefs that we have but cannot access do exist and do cause anomaly because they are beliefs whose contradictory presence in our mind causes conflict in us, even when they are both conscious, such that if they are initially conscious, tension between them might be reduced by rendering one of them inaccessible.

This raises the question of unconscious beliefs. In note 57 McLaughlin says that he hesitates to appeal to the notion of an unconscious belief as this notion 'is sometimes understood' (here, he means repressed beliefs) because an unconscious belief may be 'too deeply buried' to explain the sort of 'tension' that seems present in the minds of some self-deceivers. The typical self-deceiver is more conflict-ridden than he would be if one of his contradictory beliefs were unconscious in the sense in question (1988, pp. 60 -61).

Here he is suggesting that unconscious phenomena are less likely to motivate relevant conscious thinking than are conscious phenomena. But McLaughlin says,

> I am inclined to think that unconscious beliefs are inaccessible beliefs, though not conversely … The point to note is that inaccessible beliefs can be manifested in any way that does not involve consciously thinking that *p*. They can generate 'the tension' mentioned above in note 57 when they are accompanied by contradictory accessible beliefs. (1988, p. 61 note 60)

In saying 'though not conversely', I read McLaughlin as pointing out that although unconscious beliefs are inaccessible while they remain unconscious, an inaccessible belief can be either unconscious or conscious, as long as when it is conscious the subject is not consciously thinking about it. But if she is not consciously thinking about it then there is a sense in which it is unconscious. It is unconscious in the sense of being a standing belief that is not being thought about. This suggests that if the reason why the subject is not consciously thinking about it is that she is *avoiding* thinking about it at that time, it is an

estranged belief at that time.

This suggests that there are different degrees of estrangement. The 'Freudian' kind, which we call repression, may be severe and may last a lifetime. But there are other estrangements that are often weaker and more temporary. For example, the subject might be concentrating very hard on thinking about something quite different. Suddenly, someone reminds her of what she is trying hard not to think about. Reluctantly, she recalls it to mind and faces its implications, so becoming 'un-estranged' from it. In this sense estrangement can 'come and go', and there can be borderline cases where we feel unsure whether they are cases of self-deception or of ambivalence.[6] In note 57 more generally, McLaughlin points out that inaccessible beliefs, either conscious or unconscious, can generate tension when accompanied by a conscious, accessible but incompatible belief. When unconscious, as he points out in note 57, particularly when repressed, they do sometimes generate conflict, although in his view they are less likely to do so because they are more deeply 'buried'. I disagree with McLaughlin that the more deeply buried it is, the less likely it is that an inaccessible belief will cause inner conflict. As long as it remains in memory, it has the capacity to affect that conscious belief.[7] A fear of dogs, buried since childhood, may continue to affect one's attitude towards dogs when one is forty.

McLaughlin's position is largely but not completely compatible with Moran's. Two conscious but incompatible beliefs can cause ambivalence and thus conflict, as we saw in Chapter 2 in the case of the Catholic woman. But unconscious beliefs can also cause conflict. Unconscious beliefs that *p* can be

[6] I use 'ambivalence' to mean that the subject has two contradictory attitudes, both of which are fully conscious, but that pull her in different directions.

[7] Marshall (2000) argues that it can be active while inaccessible because it remains wilful.

inaccessible beliefs that *p*; they can be manifested in any way that does not involve consciously thinking that *p* and they can generate tension in self-deceived subjects. McLaughlin also seems to be thinking that perhaps estrangement can explain this. As long as we do not *equate* estrangement with repression, some beliefs that are inaccessible to the subject because the subject is estranged from them may be repressed beliefs, but not all. Repression, that is, will be one kind of estrangement but not the only kind. The remainder will be inaccessible for other reasons: for example, the subject may focus her attention elsewhere so that she can avoid thinking her unacceptable thought.

McLaughlin's position in 1988, well before Moran wrote *Authority and Estrangement* in 2001, is similar to Moran's in several ways. McLaughlin is claiming that inaccessible, estranged beliefs can be known unconsciously without necessarily being repressed. Repressed beliefs, he claims, are manifested via such events as dreams and slips of the tongue. I assume he would also include behavioural events in his list of events that can make repressed beliefs manifest. Conflict is present whether the contradictory beliefs are conscious or unconscious. Non-self-deceived subjects who suffer from ambivalence may suffer deeply from conflict between two contradictory courses of action which pull them powerfully in opposite directions, while both of these contradictory courses of action are conscious. But when one of two contradictory beliefs is accessible and the other is inaccessible, the beliefs may also cause conflict within the self-deceived subject[8] because unconscious beliefs can be re-activated in relevant situations, motivating the subject to maintain her self-deception.

There are differences, though, between McLaughlin's and Moran's views.

[8] McLaughlin makes this same claim about akratic subjects, discussed in Chapter 6.

Moran would argue that conflict between an unconscious and its contradictory conscious belief can be just as intense as conflict between two conscious beliefs. Moran's account also separates first-person authority from third-person authority, something that a third-person account such as McLaughlin's cannot do. McLaughlin cannot use, in his explanation, any difference between an avowal and an attributive self-report; the differences between these become invisible in a third-person, empirical account. Introducing Moran's two stances would improve Audi's and McLaughlin's accounts because if we did this we could give a clearer account of self-deception.

Moran argues that this difference in stances is most clearly exemplified in such phenomena as self-deception: self-deception is primarily a state in which 'a kind of psychological *dissociation* [my italics] gives rise to a disparity between what the self-deceiver knows, *albeit unconsciously* [my italics] and what he avows or is disposed to avow'; in such conditions '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' (2001, p. 67)[9].

I argue that when we ask why some subject is self-deceived, motivated belief is only the beginning of an answer; the questions remain i) why the bias is *there*, and ii) why the subject cannot overcome it. If the motive is *accessible* to her conscious awareness, it must be empirically possible for her either to

[9] This indicates a similarity between self-deception and akrasia. Although akrasia is about doing rather than believing, the akratic subject can be confused about what to do because she both believes and does not believe that she ought to do X or believes that she ought to do X but desires to do Y, where one of each of these pairs of beliefs or one or more aspects of it is inaccessible to her conscious awareness. For example, you may know that you ought to do X but are angry about this because you think that your having to do X is unfair. However, you cannot admit to yourself that you are angry because you need the approval of the person you are angry with. You are afraid that if you admit your anger to yourself, you might then accidentally 'show' it somehow to this person, so causing him/her to become angry with you. But I argue in Chapter 6 that not all akratic behaviour involves an estranged attitude.

overcome its force or at least to admit to it, at which point she is not self-deceived. Where she cannot do this, this is evidence that her motive is inaccessible to her at that time.

In this section, I have given my own working definition of self-deception: in self-deception the subject necessarily has two contradictory beliefs, one of which is conscious while the other is inaccessible (estranged from the subject's explicit knowledge), together with an ostensible rationalising attitude given to hide the subject's real reason from herself and others. I have supported my definition with Audi's. I have set out McLaughlin's claim, to some degree compatible with my definition, that we can maintain two contradictory beliefs as long as one of them is inaccessible and the other is accessible. I have argued that this is consistent with Moran's claim that estranged attitudes cause self-deception. I have also used McLaughlin's discussion of unconsciousness to agree with him that there are different degrees of estrangement, from severe (Freudian) to less severe and more temporary.

Next I ask whether self-deception is intentional. I argue that it is caused (at least sometimes) by a purposive act, but that the purpose of this act is not to deceive oneself, it is to avoid an unacceptable belief or intention.

Is self-deception intentional?

One traditional issue about self-deception is whether it is intentional. This is not the burning issue that it was two or three decades ago, largely because it has been replaced by deflationary approaches (such as those of Mele and Sanford).

The traditional position, that self-deception is intentional, lands us with two apparent paradoxes: the static paradox and the dynamic paradox. The static paradox has already been introduced in examples from Mele and Sanford. It is that self-deception involves two contradictory beliefs, both of which are

explicitly conscious at the same time, and that this is impossible. The dynamic paradox is that you cannot intentionally get yourself to believe something that you explicitly believe to be false.

It is easy to avoid the static paradox by stipulating that one of the pair of incompatible beliefs is inaccessibly unconscious. This solution is unacceptable to many deflationists, who do not wish to argue that self-deception necessarily involves anything unconscious.

The dynamic paradox gives us sufficient reason to reject epistemic intentionality but not prudential intentionality. However, prudential belief can be self-induced. An example of this (from McLaughlin) is the following:

> Mary wants to avoid attending a certain meeting booked for a certain date. So she deliberately writes down the wrong date in her diary, banking on the likelihood that by the time the date of the meeting comes around she will have forgotten that she did this and thus will be self-deceived into thinking that the date she wrote down is the correct date. But even if her strategy succeeds, Mary does this for prudential reasons, not for evidential reasons. This case does not explain normal cases of self-deception, where the subject's reason for denying what she unconsciously knows to be true is sincere and has not been formed by her own earlier self-manipulation. (1988, p. 37)

Mary has intentionally caused herself to become self-deceived about the meeting that she wants to avoid. There is no problem about this for a theory of epistemic self-deception; Mary's reason is purely prudential. I set this kind of self-deception aside.

McLaughlin also argues that non-prudential self-deception is not intentional. If we intentionally tell ourselves to believe what we know to be false, we will know that our new belief, if we try to adopt it, is false and will not be deceived by it. The concept of intentional self-deception seems, therefore, to be incoherent because it implies the dynamic paradox. I agree with this.

However, rejecting intentionality does not explain why it seems to us that in self-deception we deliberately avoid thinking unacceptable thoughts. If we are ashamed of our behaviour in stealing someone else's money we may tell ourselves that we borrowed it and fully intend to return it. Our insistence is sincere: we have succeeded in burying our knowledge that we stole the money deliberately. Of course, if Boyle's reflectivism is viable we do know that we have stolen the money, but we know it only unconsciously; it is safely inaccessible to our conscious awareness. This does seem to me to be purposive. But if it is, the purpose is not to deceive ourselves: it is to avoid the unacceptable. Unfortunately, however, our avoidance of the unacceptable succeeds only at the level of consciousness; unconsciously, the unacceptable remains to trouble us in the form of an estranged attitude.

Section 3: Applying Moran's account of self-knowledge to self-deception

In this section, I first set out Moran's account of self-knowledge as it applies to self-deception, explaining the inner conflict that drives this activity. To argue my case, I use the example of the mother whose daughter has learning difficulties.

On Moran's account, fully active agency and successful examples of first-person authority are synonymous: that is, a sincere self-ascription of belief, desire, intention or affective attitude has first-person authority if and only if it is a successful example of active agency. To be a successful example of active agency, a self-ascription must have the following characteristics.

It must be an avowal, sincerely self-ascribed from the first-person, deliberative stance. Moran argues that the theoretical stance must defer to the deliberative stance: 'as I conceive myself as a rational agent, my awareness of my belief is

awareness of my commitment to its truth' (2001, p. 84).

It must conform to the transparency condition. Conformity implies

> that the subject is unaware of any biasing attitude she has that might be motivating her self-ascription, implying that such biasing attitude (if any) is inaccessible to her at that time: that is, that she is estranged[10] from it;
>
> that the avowal has been formed by reflection on reasons concerning only the object of the proposition being avowed, not on any reason or reasons that are about the avower rather than about this object. To decide whether she believes that p, the avower must consider nothing but p itself; (2001, pp. 84–85)[11]

It must be at least minimally rational in the eyes of observers who are in a position to know this.

If she is estranged from her attitude, the subject cannot act as a fully active agent with respect to any intentional attitude she has that involves this attitude; the theoretical stance has usurped the deliberative stance and its requirements. Her self-ascription cannot conform to transparency because estranged attitudes are potentially active, able unconsciously to motivate conscious thinking.

This brings us to the example of the mother whose daughter has learning difficulties. The mother refuses to accept this fact because she finds it shameful.

---

[10] 'Inaccessible' throughout means 'rendered unavailable to conscious awareness *so that* the subject can avoid discovering this attitude': that is, the subject is estranged from the attitude.

[11] The subject's conformity to transparency must be settled and wholehearted. When she feels unsure about whether *p*, transparency fails because her capacity for rational self-determination has been blocked by an attitude that she cannot consider because it is inaccessible to her conscious awareness. Her belief is thus not completely self-determined. This lack of complete self-determination might be rational in cases of (say) ambivalence, where some of the evidence supports *p* and some supports ~*p*.

She ‘buries’ her belief that she feels ashamed, thus becoming estranged from it. She sincerely ‘avows’, ‘I believe that my daughter’s teacher is responsible for my daughter’s supposed learning difficulties’. This avowal is her ostensible rationalisation for her belief, ‘I believe that my daughter is not responsible for her supposed learning difficulties’. This second belief contradicts her real but estranged belief, ‘My daughter is responsible for her own learning difficulties’.

To put this more clearly, the mother’s belief A is:

‘I believe the teacher is responsible for my daughter’s learning difficulties’.

The mother’s second belief, belief B, is:

‘I believe my daughter is responsible’.

These two beliefs are contradictory. The mother cannot consciously and sincerely believe both A and B at the same time. This is why she must ‘bury’ belief B. Once B is safely inaccessible, she can confidently declare A.

Firstly, does ‘Her teacher is responsible for my daughter’s supposed learning difficulties’ conform to transparency? To do so, this self-ascription must be minimally rational in the eyes of observers who are in a position to know. Let us suppose that in this example there is ample evidence available, evidence that the mother has already been given, to show that her daughter has had learning difficulties for some years. Her mother’s self-ascription is too irrational to be acceptable to those in a position to know this so it fails the minimal rationality test. It also becomes evident to others that it is not just the evidence that the mother is considering, that there must be some other reason for her claim. They ask, ‘*Why* does she irrationally blame her daughter’s current teacher?’ The natural answer is that the mother cannot accept that her daughter has learning difficulties for reasons that she cannot admit to. But her self-ascription *is sincere*. This means that her real reason for her claim must be

inaccessible to her: that is, that she is estranged from it. She cannot admit to it *even to herself*. Thus, unknown to her at the level of conscious awareness, she has a reason for her self-ascription that is about herself (to avoid noticing her own shamefulness) rather than about her daughter. Therefore, her self-ascription does not conform to transparency. Her capacity for rational self-determination has been blocked by an attitude that she cannot consider because it is inaccessible to her. Her self-ascription is therefore not fully and actively self-determined; it is psychologically unfree.

Deflationary accounts such as Mele's reject this possibility. Mele's claim is that self-deceived subjects need not have contradictory attitudes towards their object, let alone inaccessible attitudes, which block the subjects off from being able to know or admit their real reason for their self-ascription. His examples tend to fit his theory. I demonstrated in Section 1 that all of his examples make more sense if there is an inaccessible attitude operating in them.

To give a brief example of my own: Jack believes that he loves Jill when really he does not; it is her money that he loves. His desire for her money is the real reason for his belief that he loves her, but he has rendered this reason inaccessible to his conscious awareness because he is ashamed of it.

Secondly, how can we show that an estranged attitude remains potentially active while its subject is estranged from it so that it can bias her thinking about some matter? Common sense tells us that the concept 'shameful' remains linked in this mother's memory with the concept 'learning difficulties', re-activated by later events that remind her of it. If a repressed, estranged attitude is *inactive at all times*, how can this boy's fear be re-activated in this way? Indeed, we can take this further. It might be only when he can consciously remember the event that originally caused the repression that he can begin to control his fear better and discover that dogs can be friendly. It is only when the mother whose daughter has learning difficulties can consciously

acknowledge her feeling of shame that she can begin to see that in fact, learning difficulties are not shameful at all.

If repressed attitudes did not remain potentially active while repressed, they could not cause conflict in the subject. On Moran's account (2001, pp. 83–91), conflict is motivated in two ways. Firstly, it is motivated by dissociation between the inaccessible attitude and the remainder of the subject's relevant web of beliefs. This dissociation isolates the inaccessible belief and prevents the subject from considering whether it is reasonable. Secondly, conflict is motivated by the contradiction between a subject's belief, such as 'I want a new X because the old X is past its best', and that subject's continuing, true but inaccessible belief, such as 'I want a new X because I want to have a better X than my neighbour has'. The subject knows this unconsciously. It conflicts with his new, ostensible reason for wanting a new X, because his real reason continues to be potentially active while still inaccessibly unconscious, motivating him to find his ostensible reason in order to avoid having to confront his real reason: that is, his jealousy of his neighbour's X, his feeling of inferiority at not having a better X himself.

Moran makes it clear in his account of self-knowledge that estrangement leads to anomaly via inner conflict, and that this conflict prevents the wholeheartedness in the subject's self-ascription that is required for a successful avowal (Moran 2001, pp. 76–77). He discusses such conflict in detail (2001, pp. 86–90). For example, he discusses Rey's (1988) claim that we have central explanatory attitudes (Moran's theoretical or attributive attitudes) and attitudes we can avow (Moran's deliberative attitudes), and that these do not conflict with each other because they are of different kinds. Moran points out that attributive and avowable beliefs cannot be of different kinds, because if they were they could not clash with each other. But it is only because they can clash with each other that inner conflict and thus the anomalies can occur:

> if the beliefs I express when I avow them … are simply of a different kind from the beliefs … that are the central explanatory ones, then it is completely unclear how we may see the two as clashing at all. And yet without the clash, we lose the phenomenon the distinction is supposed to help us understand. For if they are anything at all, conditions like akrasia and self-deception are some kind of conflict within the person, expressive of conflicted relations to the same thing, and this sense is lost if we see the avowed belief that *p* and the central explanatory belief that *~p* as distinct attitude types, and each all right in its own way … But avowing and reporting cannot be thus isolated from each other. (Moran 2001, p. 87)

If reporting and avowing could be isolated from each other in the way that Rey suggests, then we could not explain

> the crucial therapeutic difference between merely 'intellectual' acceptance of an interpretation, which will normally be seen as a form of resistance, and the process of working-through that leads to a fully internalised acknowledgement of some attitude which makes a felt difference to the rest of the analysand's mental life … what must be restored to the person is not just knowledge of the facts about oneself but self-knowledge that obeys the condition of transparency. (2001, p. 90)

An example that Moran gives to support his claim that a self-ascription is not a successful avowal in cases where it cannot be wholeheartedly endorsed is that of the adult analysand who becomes convinced by her analyst that she feels that during her childhood she was betrayed by another child and so can sincerely self-ascribe, 'I believe that X betrayed me' (Moran 2001, p. 85). But when she reflects *for herself* on the *object* of this relationship, the other child, she can see no reason for believing that this child has ever betrayed her. Thus she cannot endorse her belief that she believes that X betrayed her; it remains an attributive or theoretical belief. Since the transparency condition does not apply to an attributive self-ascription in this kind of situation (because she cannot normatively endorse it as true or choice-worthy), her self-ascription

cannot conform to transparency. It is not an example of active agency but rather a statement of psychological fact about the subject. If her analyst is right, her self-ascription contains an estranged attitude. If that attitude were accessible, it could tell her *why* she believes that X betrayed her. But her real reason for this belief is inaccessible. If she offers an ostensible reason (a rationalisation) for her belief that X betrayed her, she is self-deceived in believing that this is her real reason. Of course, it may be impossible for anyone else to know whether the belief she offers is her real reason, since the ostensible reason that she offers continues to affect her relevant behaviours. Self-deception is often hard to spot.

In this section I have argued that estrangement and inner conflict are the forerunners of self-deception, occurring when a subject cannot say why she believes that *p* or can give only an ostensible rationalisation as her explanation in situations where it is reasonable for others to expect that she can do this. Estrangement can be severe, or weaker and more temporary, as when the subject avoids thinking about her unacceptable attitude by focusing her attention elsewhere and by sincerely declaring an ostensible or rationalising reason to be her real reason when it is not. Estrangement produces inner conflict in the subject between her real reason for her belief and her rationalising reason for her belief. Estrangement is present in cases where the subject cannot wholeheartedly endorse her sincere belief; self-ascriptions of such a belief cannot conform to transparency and thus do not have first-person authority. This is because they are not fully self-determined and are thus psychologically unfree.

Deflationism is a third-person, theoretical approach to self-deception. As such, it ignores the fact that the first-person, practical perspective has an irreducibly different character from that of the third-person, theoretical perspective. Because it ignores this basic fact, Moran's distinction of stances and authorities

is invisible to deflationism; the practical, first-person stance and its authority, central to his account, have lost their unique character in becoming 'theoreticalised'.

Conclusion and a look forward

In this chapter I argued that when we use Moran's account of self-knowledge to increase our understanding of self-deception we can give a more reasonable explanation of self-deception than its major rivals can. Perhaps this is because Moran's position echoes what I call the commonsense position. This is that if you do not explicitly know why you believe your self-ascription, knowing this only unconsciously instead, and you invent an ostensible reason for it, you are self-deceived.

I adopted the commonsense position that the subject has two contradictory beliefs, one conscious and the other inaccessible because it is unacceptable to her and that these are in conflict. Because one of these contradictory beliefs (the unacceptable one) is inaccessible, the subject's agency is necessarily passive with respect to this belief. She can produce only an ostensible rationalisation to explain why she has her self-deceptive belief, and this rationalisation prevents her from discovering her real reason for having the belief. I found that in every case I considered, because of the passive element in her thinking, an ascriber of her self-deceptive belief suffered an estranged attitude. I concluded that the presence of an estranged reason was a necessary although not a sufficient condition for self-deception; the self-ascription of an ostensible reason was also necessary. I concluded that rival accounts are less reasonable than mine.

I began by considering two deflationary accounts of self-deception that argue that the self-deceptive process need not involve an inaccessible, estranged attitude. In the examples from Mele and Sanford that I considered, I found the

static paradox. This was not surprising. Deflationary accounts reduce avowable self-ascriptions to theoretical self-ascriptions to avoid the conclusion that the conflicting beliefs we find in self-deception can both be consciously known, at the same time, by the subject. Instead, however, they produce a static paradox that cannot be explained. It is more likely than unlikely that there is an estranged attitude motivating the subject, its presence causing the subject to fail the transparency test. I argued that where self-deception is concerned, the subject always has an inaccessible attitude —otherwise she would not be self-deceived.

Next, I briefly reviewed two accounts, those of Audi and McLaughlin, that gave some support to Moran's own. These led me to consider whether self-deception is intentional. I claimed that the subject's avoidance of a belief she did not wish to consider was sometimes intentional. However, her intention was never to deceive herself but rather to avoid the unacceptable belief. I concluded that when we use Moran's account of self-knowledge, instead of thinking about self-deception as an example of practical irrationality that might have little or nothing to do with self-knowledge, we find an explanation of self-deception that is more reasonable than the explanations given by other writers in this field.

Where akrasia is concerned, the situation seems at first glance to be quite different. Akrasia is usually considered to be identical to weakness of will, and the akratic subject is traditionally said to have free agency over her action. As I argued in Chapter 2, this has led some philosophers, such as Owens, to claim that deliberately irrational action, a kind of action that is said to be akratic, has free agency *par excellence*. However, I argued there that because, in a deliberately irrational act, the subject suffers conflict between what she knows she should do all things considered and what she does, she cannot be wholehearted about what she does. More generally, I argue that all serious

akratic acts are necessarily less than fully self-determined and that such acts are therefore necessarily psychologically unfree. 'Everyday' (non-serious) akratic acts are also psychologically unfree because their subject cannot wholeheartedly endorse her akratic self-ascription. There is a passivity operating in both self-deception and akrasia because you do not completely understand *why* you act as you do. It is not because akratic people have weak wills that they act against their own judgment, it is because they are simply not free to act according to their better judgment.

One difference between self-deception and akrasia is that the empirical *cause* of akratic action may be well known to the subject, whereas this is not the case with self-deception. For example, many alcoholics know that they are addicted to drinking alcohol. They can explain this, if asked. What they cannot do, however, is *justify* their continuing to drink. They are not sufficiently in control of why they are drinking to be able to do that. A conflict remains between their desire to stop drinking and their desire to continue drinking. Explaining it as addiction is a theoretical explanation. They need a practical explanation, and they cannot usually find one. Being told, 'You are addicted because your father died when you were three, or because your mother didn't love you … ' is just more theory. It may be true but it doesn't necessarily help in practice. For this reason, akrasia is more corrosive than self-deception. The self-deceived subject cheerfully believes that she is not self-deceived at all, whereas the akratic subject is often explicitly and painfully aware of her dilemma.

Chapter 6: Akrasia

Introduction

The word *akrasia* is the Greek term for weakness of will. (I suggest that this is misleading.) Contemporary accounts of akrasia tend to treat it as a failure of practical reasoning. A standard description of it is: to act knowingly, freely and intentionally against your better judgment through yielding to temptation or short-term pleasure or appetite in a way that conflicts with your values or principles. I reject this description for reasons I will explain. I will mostly describe it as 'acting against your own better judgment all things considered', when that judgment is explicitly known and acknowledged at the time you act.

Suppose, to use an example of Mele's, 'if, while judging it best not to eat a second piece of pie, you intentionally eat another piece, you act incontinently [akratically]—provided that your so acting is uncompelled (e.g., your desire for the pie is not irresistible)'.[75]

But if it is resistible, why do you not resist it? Are you not in control of what you are doing? Is not your action fully yours? Do you know *why* you are acting against your own explicitly known better judgment? Over time, questions like these have proved hard to answer. It seems to be a basic intuition that akratic acts are free and intentional; they seem to be acts of free and responsible agency. Why, then, do we freely and intentionally act in ways that we explicitly judge, at that very time, to be against our better judgment all things considered? This is the hard question of akrasia that I attempt to answer in this chapter.

---

[75] Mele uses this example in his entry under *akrasia* in *The Cambridge Dictionary of Philosophy* 1999, p. 16.

Stroud gives us a schematic example of how akrasia is commonly regarded today: Joseph did f rather than e, even though he was convinced that e was the better thing to do all things considered (2009, p. 1).

Stroud suggests that this is a genuinely puzzling case, one that it is hard to make sense of. She asks: why would Joseph do *f*, freely and intentionally, when he thought *e* was the better thing to do, all things considered? The hard question for the philosopher of akrasia is what the connection is between your judgment (say, that you should be doing *e*) and your action (doing *f*). Human beings normally act for reasons.

For what reason might you do *f* if you think *e* is the better thing for you to do, all things considered?

It is important to clarify at the outset a difference between akrasia and self-deception. In what are called 'everyday' cases of akrasia, it is sometimes difficult to see whether the subject is acting akratically or is self-deceived or even is not akratic at all. Suppose I know if I choose to admit it to myself that the best thing for me to do all things considered is to decline the chocolate I am being offered. But I do *not* know, consciously, explicitly and for certain, even as I accept and eat the chocolate, that I ought not to be doing so, because I have successfully avoided thinking about this. It could be said that, therefore, I do not act akratically, because *at the time* I do not know, explicitly and with conscious attention to this matter, that all things considered, I ought not to eat the chocolate. But I might be self-deceived. If I have an estranged attitude where chocolates are concerned (such as that they are bad for me) and give a rationalising explanation (such as 'one won't hurt me') when asked, I am self-deceived. I suggest that in some cases this also applies to Mele's example, quoted above, about eating the second piece of pie.

In the discussion of self-deception I argued that there is always an *unconscious*,

estranged attitude because if that attitude were accessible to the subject's conscious awareness she would not be self-deceived: the static paradox would prevent it. However, where akrasia is concerned, because of the point just made about chocolate eating, and adhering strictly to cases where the subject *does* know, consciously and explicitly even as she eats the chocolate, that she ought not to be doing so, I separate out serious from everyday cases of akrasia.

In all serious cases, I argue that the subject necessarily has an estranged attitude, where 'estranged' means 'repressed' or 'inaccessible to conscious awareness'. In everyday cases, I suggest, she may not have this kind of estranged attitude. However, if she is akratic, she will feel a different kind of estrangement, a conscious kind, that I am calling 'alienation', from doing either *a* or *~a.* Simply being in some degree of consciously unresolved conflict between doing *a* and doing *~a* is sufficient for her to lack wholeheartedness in her decision to do either. In these cases, given that she cannot conform to transparency and thus that her self-ascription is to some small extent psychologically unfree, she is mildly akratic if she acts thoughtlessly or recklessly while explicitly knowing that she ought not to be acting as she does. So why, in such cases, *does* she act as she does? My answer, in everyday matters such as the chocolate eating case, is that she does so because she knows that the consequences of doing so are too trivial to bother her. But it could be said that if *that* is why she eats the chocolate, then perhaps she does *not* know, fully, explicitly and wholeheartedly, that she ought not to eat this one chocolate at this particular time. So for everyday cases such as this, I suggest that the subject might or might not be akratic.

In Section 1, I discuss the attempts of Socrates, Aristotle, Hare, Davidson, Bratman and Mele to explain whether and how akratic acts occur. Aristotle (2011), Bratman (1979, pp. 153–171), Davidson (1980, pp. 21–42), Hare (1952)

and Socrates[76] all claim that there is a conceptual or necessary connection between believing or judging that a certain action is best and being motivated to do that thing. When we seem to break this connection, there is always some other explanation. The one that, in their different ways, these philosophers all adhere to is that all such acts involve a kind of ignorance. My own position is that akrasia does involve a kind of ignorance, but that it is a different kind from any that have been suggested by the philosophers listed above: it is the ignorance caused by estrangement or alienation, unconscious or conscious.

When we turn to Mele's account (1987, p. 94), we find no attempt to support the claim that the judgment/action connection is conceptual. Instead, Mele argues simply that when one's desire is out of line with one's evaluation of what one judges it is best to do, one may act freely and intentionally but irrationally; that is (in such cases), akratically. Mele says that 'the motivational force of a want may be out of line with the agent's evaluation of the object of the want' (1987, p. 37). One problem with this is that evaluation, since it can be overruled by the subject's desire, seems not to play any essential role; the connection seems accidental or fortuitous when it occurs. On the other hand, if desire cannot be completely divorced from evaluation and if evaluation is conceptually tied to better judgment (both of which, I believe, are the case), Mele's account does not explain how akrasia can occur.[77]

To end Section 1, I set out Tenenbaum's (1999) example of a serious case in which he argues that the subject, Joe, literally *cannot* freely do other than what he explicitly knows he ought to do, all things considered. This example provides us with some introductory evidence for my own attempt to answer

---

[76] I use Watson's 1977 discussion of Plato's Socrates here. Later I use Aristotle's discussion of Socrates' position in his *Nichomachean Ethics*.

[77] Tenenbaum (1999), for example, argues that there is a conceptual connection between motivation and evaluation in free action.

the hard question. I give it to introduce Section 2.

In Section 2, I support my account by considering examples. I begin with a case of my own. My example is of a serious case in which, I argue, postulating an estranged (inaccessible to consciousness) attitude is the only satisfactory way of explaining the subject's action. In the case I discuss, the subject suffers from a current inability that involves an unconscious motive.

I next discuss an example of akrasia that Tappolet (2007, pp. 97–120) has given us. Tappolet's article discusses the role and importance of emotions in akratic action, casting them in a more positive light than merely as non-rational causes of akrasia, and arguing that they can make akratic action intelligible and rational. Given my own emphasis on estranged attitudes, I agree that linking emotions to the rationality or irrationality of an akratic action is helpful because it is usually our emotions that we repress or, more temporarily, avoid noticing, so causing the estrangement. However, I argue that there are two aspects of Tappolet's discussion that are problematic.

Firstly, Tappolet argues that emotions can make akratic actions both intelligible and rational. They allow us 'to track reasons which we have but which we have neglected in our deliberation' (2007, p. 115). However, she includes *unconscious* emotions in her discussion, arguing that in hindsight they can make an akratic action not only intelligible but also rational. I argue that unconscious emotions cannot make the subject's action more *rational*, either at the time or in hindsight.

Secondly, because her account is deflationary, Tappolet cannot use her examples in the way she wants to. The third-person, deflationary approach obliterates the difference, basic to my account, between first-person and third-person self-ascriptions and thus between 'akratic' acts that may have first-person authority and akratic acts that have only third-person authority. I argue

that in the example I discuss, that of Emily, the subject does not have first-person authority over her action because at the time she acts she does not know why she acts as she does.

I conclude that applying my account of the anomalies to the problem of akrasia gives us a clearer explanation of both its serious and its everyday instances. *Serious* akratic acts cannot have first-person authority because they involve an estranged attitude, about which the subject is always conflicted even when she is unaware of the conflict. Thus she can neither wholeheartedly embrace nor justify her akratic act. Not every everyday akratic act need imply an estranged attitude in the sense of an attitude that the subject has repressed—she might be quite consciously alienated from both alternative actions simply by suffering unresolved conflict between what she knows she ought to do and what she does. In this conscious sense, like the sufferer of serious akrasia, she has an estranged (alienated) attitude. The everyday akratic subject's self-ascriptions therefore also have only third-person authority because they also are not fully self-determined. In everyday cases, even when no repressed attitude is involved, the subject suffers unresolved conflict between what she knows she ought to do and what she does, simply because she knows that she is doing what she wants to do instead of what she ought to do. This lack prevents her from having fully active, settled and wholehearted control over her action. Some such acts might be culpable but many will not. Some everyday akratic acts may be only mildly akratic.

So, in this chapter, I acknowledge the necessity of the conceptual connection. However, I refine this connection by arguing that no *serious* akratic act is psychologically free. My reason is that in serious cases you cannot, of necessity, *freely* do other than what you explicitly know, at the time you act, is the best thing you can do, all things considered. This implies that in serious cases, when you *do* do something other than this, you are acting to some degree unfreely;

you would act differently were you free to do so.

Section 1: The hard question of akrasia

As we saw above, all theories of akrasia must confront the hard question of the connection is between your judgment (say, that you should be doing *e*) and your action (doing *f*) in akratic action. Beginning with ancient philosophers, let us now consider in some detail attempts to answer the hard question. I argue here that none of the writers in the list can satisfactorily answer this question. To repeat my own answer, it is that in serious cases the conceptual connection between judgment and action is never broken freely; it is, however, broken unfreely. In everyday cases, I suggest, the subject may or may not have an estranged (repressed) attitude or suffer conflict between doing *a* or doing *~a* and that this conflict alienates her to some degree from doing either. In all akratic cases, the subject's action fails to have first-person authority.

Socrates and Aristotle

How do ancient philosophers explain how the connection between judgment and action appears to be breached in akrasia? For Plato's Socrates the connection is necessary and cannot be breached.[78] What goes wrong is that the subject makes the wrong judgment about what is best in her particular case. It is the judgment that is faulty; the connection between judgment and action is never broken. However, since people obviously do act against their explicitly known better judgment, it seems to be false that the connection itself is never broken. I argue that my account, using the idea of ignorance that was first suggested by ancient philosophers, solves this problem by arguing that in serious cases the connection is never broken freely. In serious cases the subject

[78] Here I use Watson's 1977 discussion of Plato's *Socrates*, p. 319.

is ignorant of *how* to act in accordance with her better judgment because her estranged attitude renders her unable consciously to know this in a fully endorsable way. In everyday cases the subject either has an estranged attitude or does not fully know her own mind even though she knows what she ought to do, all things considered: her self-knowledge is flawed.

We can understand Socrates' position better if we read Aristotle's discussion of it in his *Nichomachean Ethics* [*NE*].[79] There, Aristotle discusses Socrates' claim that 'nobody acts contrary to what is best while supposing that he is so acting; he acts instead through ignorance' (*NE* 1145b26-28).

Aristotle asks what the character of this ignorance is. We might say that when a person acts without self-restraint (loosely, akratically) he does not suppose, at the time, that he is acting as he ought to act. But the important thing to consider, Aristotle says, is whether the subject has the relevant knowledge but is not actively contemplating it, or whether he is actively contemplating it. Aristotle says 'this latter does seem to be a terrible thing, but not so if he is not actively contemplating [the knowledge he nonetheless has]' (*NE* 1145b26-28). What does Aristotle mean by '*not actively contemplating*' it? He might mean that the subject has repressed his knowledge or has at least 'put it to the back of his mind', as we say. We might say that he has 'forgotten' it, where the quotation marks around the word 'forgotten' indicate that the forgetting is somehow to the subject's advantage.[80] If this is what Aristotle means, then this is consistent with my thesis that this subject is currently estranged from this piece of his self-knowledge. But Aristotle might not mean this. He might mean simply that the subject has forgotten the relevant known truth in a completely

[79] Book 7.

[80] This does not imply that he has somehow *deliberately* 'forgotten'. See my argument against intentional self-deception in Chapter 5.

innocent way, such as because he is unwell, anxious or overworked, or because the matter is not worth worrying about. In this case, however, I would disagree with Aristotle that this action was necessarily akratic at all. I think that, in serious cases, the 'forgetting' is in the service of a biasing attitude. When it is innocent, the action is not akratic because *at that time* the subject does not know, either consciously or unconsciously, that what he does *is* against his better judgment. For example: you forget to ring your friend to tell him you cannot meet him for lunch that day. You are sincerely apologetic when he confronts you that evening. 'I just forgot', you say. If you have a reason for forgetting, relevant to the agreed meeting, a reason that is currently inaccessible to your conscious awareness, your 'forgetting' unconsciously motivates your akratic act; you cannot justify your forgetting unless some more pressing problem caused it. Perhaps you have a problematic attitude towards the discussion that you know the friend wanted to have with you at that luncheon, and this attitude has become temporarily inaccessible to you, biasing your thinking towards other things and causing you to forget your appointment. If the action is seriously akratic, there is always something relevant to it that is inaccessible to you—something that you are therefore ignorant of at the time. Where there is no inaccessible attitude, because there is no estranged attitude, consciously or unconsciously, your action is not akratic. We can check this by asking whether the subject's self-ascription of his 'forgetting' to meet for lunch conforms to the transparency condition. If it does not, your apology, 'I just forgot', although sincere, is merely a rationalisation—an excuse. It does not conform to the transparency condition because your real reason for not meeting your friend for lunch is about you, not about your friend. (You do not want to have with him the discussion he wants to have with you.)

Aristotle points out that there are two kinds of premise, universal and particular (*NE* 1146B37). This difference might be used to argue that when the universal premise gives scientific knowledge while the particular premise is about some

particular over which perception is authoritative from the start, it follows that when a conclusion arises from the conjunction of a universal premise and the particular premise, 'the soul must immediately act' (*NE* 1147a29. He gives the following example:

> if one ought to taste everything sweet, and this thing here is sweet … someone who is so capable and not prevented from doing so must at the same time necessarily also carry out this action. Whenever, then, the universal premise is present that *forbids* us from tasting sweet things, and another universal is also present, to the effect that every sweet thing is pleasant, and this thing here is sweet (and this premise is active), and by chance the relevant desire is present in us, the one premise says to avoid this: but the desire for it leads the way, for it is able to set in motion each of the parts [of the body]. (*NE* 1147a29-36, my emphasis)

Aristotle is saying here that our desire, not our knowledge, is responsible for our eating the sweet thing at such times. In particular, in cases where we are not actively contemplating our knowledge that we ought not eat the sweet thing, our action of eating it may follow automatically. This does not offend reasoned logic, because only 'the desire involved, not the opinion [knowledge], is contrary to correct reason' (*NE* 1147b 3-4).

So whenever our knowledge of some universal such as 'people should not eat sweet things' is not being actively contemplated at the time when someone offers us a chocolate (perhaps we have conveniently 'forgotten' it), we may accept the chocolate before our self-restraint has had time to come into action, or because it is also a good thing (maybe a polite thing) to do, in our opinion. This, it seems to me, is consistent with Aristotle's claim that we do not act against correct logic at those times because it is the desire, not the knowledge, that is contrary to correct reason. Desire can be a reason for doing something. It can thus be either a good reason or a bad reason for doing it; desire cannot be completely divorced from evaluation. However quickly we take the

chocolate and eat it, we would not do so if we thought it was poisonous, for example. Evaluation can occur at the speed of light.

To Aristotle, lack of active contemplation means that at this time we are *ignorant* of the universal 'people should not eat sweet things', a bit like being asleep or drunk, he says. At such times 'someone in the grip of the relevant passion either does not have this … premise or has it in such a way that his having it does not amount to his knowing it' (*NE* 1147b11-12). This 'or' reason is very close indeed to my claim that he knows it unconsciously and is thus temporarily estranged from it at that time.

Aristotle goes on to conclude that what Socrates argued is correct:

> And because the ultimate term is not universal and seems not to be knowable as the universal is knowable, it seems also that what Socrates was seeking turns out to be the case. For it is not when science in the authoritative sense seems to be present that the experience of the lack of self-restraint occurs, nor is it this science that is dragged around on account of passion, but rather that knowledge which is bound up with perception. (*NE* 1147b13-19)

Let us equate what we call acting akratically with what Aristotle calls acting with lack of self-restraint. We can now say that according to Aristotle, Socrates was right: akratic action can occur without contradicting reason because at the time we act akratically we can be ignorant of what we know (such as by having 'forgotten' it). Against Socrates, however, I would argue that there are many cases where we do know that eating the sweet thing is against our better judgment even as we eat it, and that Socrates' position cannot explain these cases.

The main difference in general between Socrates and Aristotle is that whereas Socrates believes akrasia is caused by mistaken judgment, Aristotle is more nuanced, saying that although the subject knows what is best in some sense,

he does not use his knowledge at the moment of decision—perhaps because he is distracted by strong contrary desires. I disagree that in serious akrasia, at least, because our ignorance is similar to our being asleep, there is never any conflict operating when we act akratically: conflict, I argue, is basic to serious akratic action. It is because conflict renders the subject unable fully to accept her decision to do *a* that her action, regardless of whether she does *a* or ~*a*, is psychologically unfree. It does not matter whether she can think clearly or not, unless her thinking can resolve the conflict. If one of the conflicting premises or the reason for this premise or its experiential immediacy is inaccessible, she cannot resolve the conflict.

Aristotle's ignorance thesis is different from my explanation in terms of estrangement, not only because it does not mention the presence of anything unconscious or of any conflict but also because it does not suppose that the 'forgetting' of what would be best to do must be in the service of avoiding what the subject does not want to remember at that time. It claims that we might just innocently forget that we know this. But if we forget *innocently,* at the time we do *b*, that we ought to be doing *a* instead, we are *not* explicitly aware that we ought to be doing *a* instead of *b*, and thus our action is not akratic.

Hare

Hare rejects Socrates' claim that the akratic subject is ignorant concerning how they should act. Instead, Hare claims that the akratic subject explicitly knows there is another possible course of action available to him that is better than the one he is following. But Hare also argues that it is impossible for an agent to do what he does if he genuinely and in the fullest sense judges that it would be better for him to do something else instead. This, Hare argues, is because evaluative judgments are action-guiding, and thus provide, in themselves, a 'Let me do *a*' answer that serves as a command or an imperative. Therefore,

sincerely assenting to the evaluative judgment 'I ought to do *a*' necessarily causes *doing a'* (Hare, p. 79). So how can a subject who sincerely assents to 'I ought to do *a'* do ~*a* instead?

Hare's answer is that the subject can do this when he either fails to realise (that is, does not know, is ignorant) that the general imperative 'I ought to do *a*' applies to him in that particular case, or he is physically or psychologically unfree to do *a.* If he thinks that 'I ought to do *a*' does not apply to him, he may be using the evaluative term 'ought' to mean only that most people would say he ought to do *a*. (1952, p. 120). Hare's view is that if he does *b*, it *follows* from his having done *b* that he judged that *b* was the best thing to do among the options available to him at that time, even if he *thinks* that *b* was not the best thing to do. If he is psychologically unfree to do what he knows he ought to do, then according to my own position, i) he is ignorant of some element in this matter that he could otherwise use in order to decide *correctly* what to do; and ii) in serious cases, this element is an inaccessible (estranged) attitude preventing him from acting as he knows he should.

Hare's explanation is the closest to mine of the accounts that I am considering. However, I do not include 'physically' as he does because physical incapacity does not imply an estranged mental attitude. Hare does not recognize that the anomalies of self-knowledge are inextricably tied to akrasia, because he does not see how such self-estrangement is required for genuine akrasia.[81]

[81] Also, Hare's view is that agents who make exceptions for themselves will be focused on some morally irrelevant factor. (A morally irrelevant factor might be, e.g., eye colour or the day of the week.) Akrasia, in his view of it, is always about *moral* judgments, which he takes to be universal. Universalisability is a conceptual requirement on moral judgment—I am not making a genuine moral judgment unless it applies to me in the relevant circumstances.

Davidson

Like Hare, Davidson rejects the Socratic claim that akrasia is explained by ignorance or some kind of illusion. He goes further, rejecting Hare's claim that ignorance must be involved somewhere. Instead, Davidson claims that akratic acts are possible because instead of acting in accordance with her own better 'all things considered' judgment, the subject can act according to her own unconditional or 'all-out' judgment. The all things considered judgment, Davidson believes, is only a prima facie evaluative judgment. It is relational, and does not involve a commitment to the superiority of the option in question; it does not imply that the all things considered judgment is also the best thing to do *sans phrase* (all out), unconditionally. Relational judgments do not tell us what is better simpliciter, but what is better in light of all the reasons the agent considers relevant. It is conditional in form. This means that to choose not to do what is better all things considered is not to make a logical mistake. However, since the all things considered judgment covers all the evidence the subject can know about, to choose *not* to do what is better all things considered is to act irrationally.

Davidson's view thus succeeds in explaining how akratic action is conceptually possible on the ground that it is possible to act irrationally in such cases. Where everyday akrasia is concerned, I agree with him that an akratic act can be irrational. However, in serious cases, I argue, Davidson's position cannot be sustained because in serious cases the subject desperately wants to do what he knows is best as far as he can tell, and yet he still does not do this. How can he not do *a*, when he desperately needs to do *a*, if he is *able* to do *a* and believes that doing *a* is best as far as he can tell? Davidson's position cannot answer this question. My answer is that he is psychologically *unable* to do *a* freely. In such cases, I argue that when he does *~a* instead, he acts unfreely.

Bratman

Bratman argues that free and intentional akratic action is possible. His argument follows from the conclusion we have just reached, that people do not always desire what they think is best for them all things considered, acting instead on a desire that they consciously believe, at that time, is not best for them all things considered. This may be irrational. However, Bratman argues, the irrational process Sam follows is nonetheless a reasoning process. His view seems to explicate Davidson's claim that an agent may act on a subset of reasons that support the action all-out, rather than on all of their reasons, all things considered.

I quote Bratman's case of Sam as follows:

> Suppose Sam is sitting by a bottle of wine. He knows that if he drinks the wine he will suffer a bad headache in the morning and further that, given the late hour, he must go right to sleep if he is to be fully rested for an important job he must do the next day. Still, he knows that drinking the wine would be quite pleasant and would, temporarily at least, help relieve his present depression.
>
> Sam thinks both that in certain respects his drinking would be, prima facie, best, and that in certain other respects his abstaining would be, prima facie, best. Weighing these conflicting considerations he concludes that it would be best to abstain, rather than drink. 'Still,' he thinks to himself while focusing his attention on his reasons for drinking, 'the wine does look very good, and it would surely lift me out of my depression.' He then reaches for the wine, pours it into a glass and proceeds to drink it.
>
> Sam drinks the wine for a pair of reasons: the pleasure of drinking and the temporary relief of his depression. He does not drink it impulsively or impetuously; he drinks it, rather, deliberately and with knowledge of his reasons both for and against drinking it. Finally, Sam is not compelled so to act. He could put the glass down and go to sleep and he knows this. Sam's case is, it seems, one of free, deliberate, and purposive—that is, full-blown—action contrary to the

agent's best judgment. It seems to be a case of weak-willed action.

Sam's action must correspond to an appropriate practical conclusion. But the only natural candidate for such a practical conclusion seems to be just his conclusion about what it would be best to do. But this is the conclusion contrary to which he must act if his case is to be one of weak-willed action; for it is his acceptance of this conclusion which constitutes his 'best judgment' concerning what to do. (1979, p. 156)

At this point Bratman claims that although Sam can see and theoretically accept all the reasons for not drinking that there are, his action does not reflect his 'best' judgment.

Bratman first rules out what he calls the extreme externalist response, which finds no essential relation between desiring to do something and valuing doing it. It sees judgments about what would be best as playing no special role in the practical reasoning underlying full-blown action. This saves the empirical possibility of akratic action by severing the connection between evaluation and practical reasoning; we do not consider our values when we act. But this is counter-intuitive because it supposes that there is nothing perplexing about acting contrary to one's better judgment. Rejecting this extreme externalist response, Bratman argues that to allow for weak-willed action we must drive a wedge between best judgment and practical conclusion, and that to do this we must give a different account of what a practical conclusion is.

The account Bratman proposes breaks the internalist[82], evaluative tie between theoretical premises on the one hand and a practical conclusion on the other hand. The conclusion is irrational, but it still follows a reasoning process.

[82] I use *internalist* here to mean the idea that evaluative judgments have an internal or necessary connection to motivation and action that purely descriptive judgments do not have.

Therefore, it shows us how akratic action can occur as part of a judgment based on a deliberative process that uses reasons: the explanation is that the subject's use of these reasons is irrational.

Bratman sets out Sam's argument as follows:

1. Given that drinking would be pleasant and relieve my depression, my drinking would be, *prima facie*, best.

2. Given that abstaining would allow me to avoid a headache and be rested for tomorrow's job, my abstaining would be, *prima facie*, best.

3. Given the considerations cited in both 1 and 2, my abstaining would be, *prima facie*, best.

4. Abstaining would be best.

So far, Sam has reasoned logically. His irrationality lies in nonetheless reasoning from 1 to

5. I shall drink.

Bratman argues that although it is impossible for Sam to disagree with his own evaluation that his grounds for abstaining are stronger, it is possible for his reasoning, in practice, to be irrational. For example, he can leap, irrationally, from 1 to 5. Thus this example, like Aristotle's 'forgetting' cases, shows that akratic behaviour can occur. Bratman cites such factors as Sam's depression, the strength of his desire to drink, and 'the way his depression dulled his appreciation of his reasons for abstaining' (1979, p. 170) as possible explanations for his irrational reasoning. I find this puzzling, because Bratman also tells us that Sam does not feel impetuous or impulsive and that 'he drinks … rather, deliberately and with knowledge of his reasons both for and against drinking'. This seems unlikely, given that Sam is depressed. Bratman concludes

that weak-willed action is possible, a conclusion that supports the common position that we can be weak-willed when we act against our explicitly known better judgment.

There are many possible ways in which we can think irrationally about what to do, including leaving out certain pieces of evidence and favouring others for reasons caused by (say) an emotion attached to some belief or desire that is biasing our thinking. As Mele argues[83], the extra force of our desire to do *a* can overcome and distort our evaluative reasoning process by biasing it. However, Bratman's conclusion also suffers from a problem: although Sam *can* employ a reasoning process leading him to the conclusion that he can drink that night, *why* does he employ that particular reasoning process while being explicitly aware of the excellent reasons he has for abstaining? Must he have an inaccessible reason for doing so? For example, might he need to drink to dull his aggression, which if not dulled might result in his attacking someone? This will depend on how serious Sam's drinking problem is. If it is very serious, then I argue that something like this might be the case. If it is less serious, then surely the most likely explanation is similar to the explanation of why I take the chocolate I am offered—it won't kill me, and I like chocolate, and it will please my hostess and me. On this explanation, Sam decides to drink to satisfy his most urgent reason: to relieve his depression. This is his reason simpliciter or *sans phrase*.

If the force of your desire to drink (a force fuelled in Sam by the pain of his depression) *overcomes* your reasoning power, this seems to imply that you are at that point *unable* rather than too weak-willed to abstain. In Bratman's example it may imply an alienated attitude. Sam knows explicitly and seemingly with great clarity why he should not drink that night. If we put that knowledge

[83] I discuss Mele next.

against the force of his depression, we can see that he suffers conflict between his desire to drink on the one hand and, on the other hand, all the reasons he knows of against drinking. He is fully conscious of both. His conflict between these alone is reason enough for concluding that Sam does not fully know his own mind about whether to drink that night. We do not need to hypothesise that he also unconsciously suffers from aggression. His desire to drink and his desire not to drink are both to some degree alien to (estranged from) him. He can neither drink, nor not drink wholeheartedly. The case of Sam thus implies that the action that follows is psychologically unfree to some extent, although perhaps only to a small extent, because whatever he does, his decision to do it is not fully *his*. Its passive element restricts his capacity to make his decision wholeheartedly.

This case supports my claim that serious akratic action always involves an estranged or alienated attitude in conflict with what the subject knows would be best for him to do. Sam's drinking that night is unfree to some degree because his alienation from both drinking and not drinking restricts his capacity to make a wholehearted choice between them. In this specific sense of 'unfree' he must choose drinking because he is psychologically *unable* not to. Bratman, like Davidson, shows us in his case of Sam that akratic action can occur when the subject reasons irrationally. But he cannot explain *why* Sam would draw an irrational conclusion from such excellent, explicitly known reasons against this conclusion. My claim that serious akratic action is psychologically unfree, on the other hand, can explain why he drinks that night. He drinks because he must.

At this point my position remains that because akratic subjects suffer an estranged attitude in serious cases, serious akratic action is i) psychologically unfree and ii) fails to have first-person authority because the subject cannot endorse it.

Mele

Mele explains akratic acts as acts where the strength of motivations does not correspond to the strength of evaluations: what the agent judges to be best to do is not what he is most motivated to do. As already mentioned, he tells us that 'the motivational force of a want may be out of line with the agent's evaluation of the object of the want' (1987, p. 37). He thus rejects accounts, like Davidson's, that explain akrasia as the product of two kinds of judgments. Mele argues that, nonetheless, the agent can act akratically but still freely on what he is most motivated to do; his akratic action is not compelled. Mele points out that strict akratic actions (actions that are akratic but not compelled) can be explained in terms of the perceived proximity of the rewards promised by the incontinent action, the agent's motivational level, his failure at self-control and his attentional condition (1987, p. 92). But in pointing to a failure of attention, for example, Mele is virtually acknowledging that something such as anxiety might be causing this lack of attention, and to move from this to the claim that the subject might be estranged, either from this anxiety or from the mental state it is preventing him from considering, is a small step.

Mele denies the Aristotelian view that when we act akratically we have always 'forgotten' what we know at other times to be the better thing to do. In other words, there are times when our judgment is correct, and we know, with full conscious awareness at the time we act, what it is. However, we may not *want* to do what we know would be overall best for us to do according to our own values; we may desire to do the opposite, and this desire might be stronger than the desire to do what we judge is best for us to do even when our judgment about this continues to operate with our full conscious awareness of it. Mele argues this way with the following example:

> Consider … a man who, on the basis of an assessment of reasons for and against his continuing to smoke, has judged it best to kick the habit and has resolved to

> do so. Such a man might, on occasion, have a desire to smoke whose motivational force does not accord with his assessment of the merits of satisfying that desire. In some such cases, the agent's judging it best not to smoke a cigarette might, owing partly to the strength of his competing desires, fail to issue in a corresponding intention; and, succumbing to temptation, he might intend instead to smoke and proceed to execute that intention. (1987, pp. 44–45)

That is, the judgment based on this smoker's consciously aware knowledge that he ought not to smoke might simply fail to prevent him from smoking. We often find ourselves in this kind of situation. Aristotle's and Socrates' accounts both fail to explain it; Hare's account lists it under psychological reasons for the subject's inability to do what he judges is best to do. Also, there are many cases where you do not fully believe that smoking will kill you in the end, and in those cases you may not be acting akratically. You are not acting against your better judgment at such times because you do not really believe that smoking cigarettes *will* kill you. Perhaps your father smoked until he died at age 102. Your act will be akratic, however, if i) your lack of belief is a rationalisation designed to hide the fact that unconsciously, you do believe that smoking will kill you; or ii) you suffer conflict between consciously thinking that it will and consciously hoping that it won't. People are often unsure about following this or that prescription given by an expert, and we are all unsure to some degree about anything that may happen in the future. When we act as we do think is best, all things considered, we are usually confident and hopeful but never completely certain, simply because the future is never completely certain. Politicians often have great difficulty in deciding to spend large amounts of money preventing what might never happen anyway.

Mele's account also faces a different difficulty: that if he is right and our desire is independent of our best judgment, then even in the face of our continuing conscious awareness that our best judgment forbids our doing *a*, our motive for doing *~a*, rather than our evaluative judgment, is *in control* of our action.

This does not seem very different from compulsion, unless we can say that we somehow *let* our motives overrule our best judgment. And why would we do that, if we could prevent it? Mele cannot explain this. I believe that in serious cases we need the concepts of estrangement and of unfree akratic acts to explain it.

Tenenbaum, in discussing Mele's view of akrasia, reports Mele's case of the agent who thinks for moral reasons that it is slightly better not to enter strip-tease clubs than to enter them, arguing that this agent could choose to enter them without contradiction and thus to enter them permissibly.[84] Tenenbaum (1999, p. 887) argues that if doing A (entering the strip-tease club) is permissible, A is not immoral and thus, presumably, also not irrational. He asks in what sense, therefore, is doing A akratic? He then says that Mele cannot use his example of doing something less than the best thing to do to show that this is not self-contradictory where the act in his example is *overwhelmingly* less good, not just slightly less good, than the best thing he could do. This, Tenenbaum says, is because it is much harder for Mele to make his point in this latter case. I agree. My separating everyday cases of akrasia from serious cases is based on this point.

Tenenbaum also discusses Mele's claim that the agent who failed to exercise self-control can still be considered free if she overlooked that it was possible to exercise self-control or if she misjudged the amount of self-control necessary and thus did not realise that it was in her power to exercise self-control. Tenenbaum points out that this would account for only a small subset of akratic acts that are said to be performed freely. He also raises the question whether the agent who fails to exercise self-control might be compelled to act

[84] Mele 1987, p. 28; reported in Tenenbaum 1999, p. 886. Tenenbaum footnotes that he is unsure of what 'permissibly' means here.

as she does. He compares this possibility with someone held at gunpoint who does not realise that he can overpower his aggressor. Failing to exercise self-control because you do not realise that you can successfully exercise such control 'should not be seen as acting free from internal compulsion' (1999, p. 889; discussing Mele 1987, p. 25).

I end this discussion of Mele's account of akrasia with Tenenbaum's example of a serious case, that of Joe and his sister. This case, I argue, supports my position.

> Let us assume that Joe lives in a dictatorship, and the dictators are after his sister who is the head of an underground guerrilla group, and Joe knows that they will show no mercy towards his sister. The government offers Joe a million dollars to turn her in and he refuses, because he finds it abominable to trade his sister's life for a few bourgeois comforts. The government explains to Joe that the offer will still be standing for 48 hours, and that they will bring to his home tonight a suitcase full of cash, hoping he'll have changed his mind by then. They'll open it in front of him, grin and wait; and Joe knows he cannot resist such a vivid display of hard currency. Fortunately, Joe can avoid facing this irresistible temptation by just putting a sign on the door that says: 'Joe does not live here anymore'. Not being very bright, the government agents will turn around and never come back. (1999, p. 887)

Tenenbaum argues that as long as Joe is minimally logically competent, he must conclude that it is better to post the sign than not to post the sign (1999, p. 888). Of course, he can fail to act on that decision when confronted with the cash. But in that case he acts unfreely because he has an estranged attitude: it turns out that unconsciously, he prefers to have a million dollars than to save his sister. If this attitude were not estranged, he would not have agonised for so long about what to do—he would have just taken the cash.

I think that Tenenbaum's point is right and important. It suggests that the less

serious an akratic example is, the more likely it is that the subject can be rebellious or indulgent or foolish about it while being only mildly akratic or even not akratic at all. For example, Owens' slumper, who watches TV instead of making lunch, becomes much 'less akratic' when measured against the case of Joe and his sister. In mild cases the subject is often in control of her action. When she is confident of being in control, and when she is right about this, then she might not really think that she *is* going against her better judgment, even if she knows that hers is a minority view. Thus in these cases she might not be acting akratically. But in serious cases, she is not in control of her action; the consequences of acting against her better judgment are too dire for her to do so freely.[85]

This concludes my survey of Socrates, Aristotle, Hare, Davidson, Bratman, Mele and Tenenbaum. I have argued that none of these writers can answer the hard question. My claim is that if serious akratic action is unfree then the conceptual connection between judgment and free action remains. The only refinement of it that we need is to claim that in serious cases, at least, it is not broken freely.

Section 2: Some examples

In this section I first offer a serious case of my own to show that such cases, at least, can be better explained via the claim that the subject suffers an estranged attitude. I then defend my position more generally by discussing Tappolet's 'Emotions and the Intelligibility of Akratic Action' (2007). Tappolet argues that emotions are central to akratic action, and I agree with this. However, she also argues that in hindsight, emotions that were unconscious at the time can make an action more rational *at the time it was performed*. I argue that they can

[85] My conclusions about freedom do not imply any particular conclusions about responsibility, but I cannot pursue that issue here.

make it more intelligible but not more rational.

The role of the estranged attitude

In *Authority and Estrangement*, Moran says

> 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 behaviour 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. (2001, pp. 127–128)

This gives us his position on akrasia. In the serious cases of it that interest him, akratic action is a kind of obsessional or compelled belief or action. The compulsion Moran speaks of, however, is not the compulsion that an addiction causes (although addiction is often very reasonably cited as a cause of akratic action) but rather the compulsion of the estranged attitude that is blocked off from its relevant web of beliefs and has become isolated in the subject's mind. This isolation prevents the subject from believing that *p*, thereby compelling her, if asked, sincerely to deny her unconsciously known truth that *p*, by self-ascribing its opposite, 'I believe that *~p*'. To distinguish this from other kinds of compulsion such as addiction, I shall refer to it here as the subject's compulsive denial of her unacceptable truth.

I agree with Moran that akratic action is puzzling. I take seriously Mele's point about emotions [86] and I ask how we can avoid the dilemma previously mentioned: that is, that there seems to be a contradiction between acting freely on the one hand and acting under emotional force on the other hand. To answer this question I first offer an example of my own that I think most writers

[86] Mele's point is that 'the motivational force of an agent's wants may be out of line with his evaluative assessment of the wanted items' (1987, p. 43). Emotions motivate us to pursue our desires.

would agree is akratic. It is a serious example because the subject is desperate. I argue that treating it as caused by a lack of willpower is not the best way of explaining it, and I offer my own explanation, that the subject has an estranged attitude.

I begin with my definition of an estranged attitude, as follows:

When the subject suffers an estranged attitude, she necessarily cannot know why she believes that *p* or intends to do *a* in *all* of the following ways: consciously, immediately, wholeheartedly and in a settled way. She might know this in some of these ways but not in all of them. It follows that

- Her knowledge of this estranged attitude can be either conscious or unconscious.
- If it is unconscious, it is directly inaccessible to her; at the level of conscious awareness she is ignorant of it.

If it is explicitly conscious, then either

- the reason she has it is inaccessible to her (and so she is ignorant of this reason) or
- she knows this reason only theoretically—its phenomenological immediacy is inaccessible to her (and so she is ignorant of how this immediacy would feel).[87]

In all of these situations, because of her estranged attitude, the akratic subject is ignorant of why she does what she explicitly knows is against her better judgment. Her action is not fully *hers*; she is not fully in control of it. Because of this ignorance, her self-ascription, although *agential*, is not fully and actively

[87] This is a bit like Socrates' and Aristotle's claims that sensory experience can sometimes convince us to do *a* more thoroughly than a theoretical reason can convince us to do ~*a*.

self-determined. It is therefore psychologically unfree, and thus it has the authority only of third-person agency.

Here now is my serious case.

John the gambler

John is a middle-aged man, married and with a teenage family. He has gambled for about twenty years and is now desperate to stop. He has already had to sell the family home to pay his gambling debts, and his wife is about to leave him. John knows that he should stop gambling and desperately wants to. But for some reason that he cannot explain, he continues to gamble. If he could bring to conscious awareness his real reason for gambling, he would be able, after some time and with much difficulty, to stop gambling.

How can we explain this situation? Why would John the gambler *not* stop gambling, given his desperate family situation, if he *were able* to do so? This is a man who *sincerely* self-ascribes, 'I want, more than anything, to stop gambling. Gambling is destroying my family. I don't know why I can't stop'. Notice that he says, 'I *can't* stop' and means it. Clearly, he *feels*, rightly or wrongly, genuinely *unable* to stop gambling.

If we allow that in serious akratic situations the subject is unable rather than too weak-willed to resist acting as he does, an alternative explanation for John's predicament becomes available. The inability that John suffers is not the inability to resist temptation or overwhelming desire. It is nothing, in fact, that he is aware of. He is unable to stop gambling because he has a stronger reason for wanting to continue, a reason of which he is unconscious because he has repressed it or has avoided thinking about it in some other way, such as by 'forgetting' it. We can explain serious cases of akrasia such as John's, I suggest, only if we introduce into our discussion of this anomaly the notion that there are unconscious forces operating in us and that these forces can sometimes

explain such cases. This is to say that the concept of weakness of will cannot always satisfactorily explain cases as serious as John's. Let us see, therefore, if we can explain John's predicament in terms of his having an unconscious attitude, an attitude that he *does have* but that is inaccessible to his conscious awareness (he knows only tacitly, unconsciously, that he has it).

Let us suppose that John's father was a sailor, and away from home nine months of each year. Whenever he headed off, he would give John twenty dollars. He would say, 'This is so that you can look after your mother while I'm away'. He began to do this when John was eight, and he died when John was twelve. How is this new information relevant to John's current situation, thirty years later?

As an eight- to twelve-year-old child, John learned from his father about the connection between money and looking after his family. He still believes, unconsciously, 'I need money to look after my family'. This belief has remained with him, albeit unconsciously, because an emotion is attached to it. This is that unconsciously, John still longs for money from his deceased father because money, while he was getting it as a child, always made him feel worthwhile. Now, as an adult, John knows consciously that the money he needs, in order to feel worthwhile, comes from his pay packet. But he still believes unconsciously that to feel worthwhile he must please his father (now identified with himself) by looking after his mother (now identified with his wife and children). His day job scarcely brings in enough money for his family to survive. To feel worthwhile, he needs more, and he turns to gambling to get it. But gambling loses him more money than he wins. The more money he loses by gambling, the more of a failure he feels and the more desperate he becomes to win next time. He feels that only by having 'the big win' can he justify his father's faith in him to look after his loved ones. John's predicament is serious. He cannot access his real attitude towards money and its connection in his

mind with looking after his family. But it is there, feeding his gambling habit, and it is destroying him.

But even given the foregoing, why does John's not knowing, explicitly, consciously and for himself, *why* he gambles *interfere with* his acting on his knowledge *that* he must give up gambling? There are two reasons why it interferes with how he acts. On my account, i) John's knowledge that he should give up gambling is theoretical only, and is thus susceptible to conflict and hesitation. (People do have big wins. Why shouldn't he?) More importantly, ii) however strongly he tries to stop gambling, there is always something that he actually *wants to do more*. This is to please his deceased father and thus himself by looking after his loved ones so that he can feel worthwhile as a husband, a father and a man. This inaccessible attitude is stronger than his conscious reasons for not gambling. It continues to motivate him to gamble and it trumps his conscious reasons every time.

The example of John suggests that in serious cases, the solution to the problem of akrasia begins with the distinction between first- and third-person stances. When the subject does not know his own mind about why he gambles, he is not fully in control of his action. John knows *that* doing *a* is not in his best interests, but he does not know, immediately and for himself, *why* it is not. (The evidence is that it is not, but the fact that he has not had his big win so far does not prove *to him* that he will not have it next time.) To know immediately and for himself why doing *a* is not in his best interests, he must explicitly and consciously associate his current longing to gamble with his still active, childhood longing to please his father by looking after his mother. If this real motive becomes consciously accessible to him, John will realise that it does not apply to his adult situation. But until then, he remains compelled to gamble by his own inaccessible desire for money. His self-ascriptions have only third-person authority. Deflationary accounts of akrasia cannot differentiate

between the first- and the third-person perspectives. But on my first-person account we can distinguish between these.

On my account, self-ascriptions in serious akratic cases do not have first-person authority for a very simple reason: conflict prevents wholeheartedness.[88] Whenever this situation occurs, the subject has a relevant, inaccessible (estranged) attitude. She may have a physical compulsion or may suffer depression. But in all these cases, she suffers conflict. This conflict renders her *unable*, at that time, fully to endorse her relevant self-ascription; e.g., 'I intend to stop *a-ing* tomorrow', because she does not fully know her own mind on this matter. For this reason, her first-person authority over her 'avowal' fails her.

When the subject's behaviour has fully active agency, she is in control of what she is doing because she knows her own mind concerning why she is doing it and can justify it.[89] In these cases, I argue that her action has first-person authority, and that for this reason it is not akratic. This is not like the kind of case discussed in Chapter 2 as a deliberate, intentional but irrational action. Owens claims that such an action is freely and intentionally made, and is an example of free agency *par excellence*. I argue that a subject who performs a deliberate, intentional but irrational action is always conflicted because she knows that what she wants to do is irrational. Thus, her first-person 'avowal' fails. It has the force only of a third-person attribution.

My account is better as a way of deciding what the akratic subject can and cannot do because it enables us to clarify akrasia by showing that although the

---

[88] The converse, that wherever wholeheartedness is lacking the action is akratic, does not follow. In cases of ambivalence, for example, the subject may be conflicted between two alternatives. Although she does not have first-person authority over whichever alternative she chooses, this need not mean that she is being 'weak-willed' about it.

[89] There might be borderline cases that seem not fully to fit either category.

subject *seems* to avow her intention to act against her own, explicitly known, better judgment, she tells us only what she intends to do, not what she will be able to do.

My example of John the gambler is meant to explain why it is that someone can continue to act akratically in some particular area of his life even though he desperately wants to stop acting this way, rather than because he is physically addicted to what he is doing or is overwhelmed by desire or temptation. John *wants* to stop gambling.

I turn now to one of Tappolet's cases, found in her 'Emotions and the Intelligibility of Akratic Action'. This case, that of Emily, discusses the role and importance of emotions in akratic action, casting them in a more positive light than merely as non-rational causes of akrasia, and arguing that they can make akratic action intelligible. Given my own emphasis on estranged attitudes, I must agree that linking emotions to the rationality or irrationality of an akratic action is helpful, because it is emotions, attitudes that impact on us emotionally, that we may repress, so causing estrangement. However, there are two aspects of Tappolet's discussion that are problematic.

Firstly, her account is deflationary; the discussion is from the third-person perspective. Because of this, she cannot use her examples in the way she wants to because the third-person, deflationary approach obliterates the difference, basic to my account, between first-person and third-person self-ascriptions and thus between 'akratic' acts that have successful first-person authority and akratic acts in which first-person authority fails. I argue that in Tappolet's example of Emily, the subject's avowal fails because at the time she acts, she does not know her own mind about *why* she acts as she does. She suffers a degree of estrangement concerning both alternatives.

Secondly, in arguing that emotions can make akratic actions both intelligible

and rational, Tappolet includes *unconscious* emotions, arguing that in hindsight they can make an akratic action not only intelligible but also rational. I agree that although unconscious emotions cannot make an akratic action intelligible to the subject at the time she acts, we may be able to say in hindsight that an emotion did make some particular action intelligible. However, I argue that on my first-person account, unconscious emotions cannot make the subject's action *rational*, whereas Tappolet argues that they can. My problem with the case of Emily[90] is that the conclusions she draws from it are wrong. Here is the example.

Emily

Emily's best judgment has always told her that she should pursue a Ph.D. in chemistry. But as she proceeds through a graduate program, she starts feeling restless, sad, and ill motivated to stick to her studies. These feelings are triggered by a variety of factors which, let us suppose, are good reasons for her, given her beliefs and desires, not to be in the program. The kind of research that she is expected to do, for example, does not allow her to fully exercise her talents; she does not possess some of the talents that the program requires, and the people who seem most happy in the program are very different from her in their general preferences and character. All these factors she notices and registers, but they are also something that she ignores when she deliberates about the rightness of her choice of vocation: like most of us, she tends to find it hard, even threatening, to take leave of a long-held conviction and to admit to herself that the evidence is against it. Still, when she deliberates, she concludes that her feelings are senseless and groundless. One day, on an impulse, propelled exclusively by her feelings, she quits the program, calling herself lazy and irrational but also experiencing a (to her)

[90] Arpaly 2000, p. 504.

inexplicable sense of relief. Years later, happily working elsewhere, she suddenly sees the reasons for her bad feelings of old, cites them as the reasons for her quitting, and regards as irrational not her quitting but rather the fact that she held on to her conviction that the program was right for her for as long as she did.

> Emily, I would like to argue, acts far more rationally in leaving the program than she would in staying in the program, not simply because she has good reasons to leave the program but also because she acts for these good reasons. (Arpaly 2000, p. 504)

But for what good reasons did Emily leave the program? The example tells us that she left while feeling restless, sad and ill motivated to continue. There is good cause for these feelings: she is not allowed fully to exercise her talents, she does not have some talents that the program requires and the other people in the program are very different from her—she does not easily warm to them. She concludes that her feelings are senseless and groundless. But it is on the basis of these negative feelings (feeling lazy, irrational, restless, sad and ill-motivated) that she leaves the program. Arpaly says that she leaves for her good reasons. But surely this is incorrect. She has set aside her good reasons as senseless and groundless. She knows that she has these reasons for wanting to leave the program but she does not leave it explicitly for these reasons.

Tappolet says:

> Emily's feelings of restlessness, sadness and ill-motivation, which cause her to abandon a chemistry PhD against her better judgement, are in fact responses to the fact that the programme is ill-suited for her. These feelings can be seen as responses to factors which, given Emily's beliefs and desires, are in fact good reasons for her to abandon the programme. Thus her emotional states not only make her decision to abandon intelligible, but indeed lead her to act in a more rational way, something which she recognizes later. (2007, p. 116)

Yes, her restlessness, sadness and ill-motivation do cause her to leave the program. But these causes are not reasons for her at the time she does that. From a first-person point of view, Emily does not use her emotional states as a reason for abandoning the program. She pulls out in despair, not because she thinks the program might be unsuitable for her. Emily does not know, at the time, why she feels restless, sad and ill-motivated to pursue her program. She realises the relevance of these feelings only later. Only later can she see *in hindsight* that she felt restless, sad and ill-motivated *because* the program was unsuited to her. Only in hindsight can she see that these feelings were a good reason to pull out. She feels confused and defeated—a failure—when she pulls out. Her emotional state does not make her decision to abandon her program intelligible even to her, let alone rational, at that time. Later, looking back, Tappolet claims, Emily realises that her unconscious emotional states 'led her to act in a more rational way'. But Emily, in looking back, is considering her action from her first-person perspective. From that perspective, she cannot say that she acted more rationally because of her unconscious emotions. She can see that those emotions explain causally, in a third-person way, why she acted as she did, but not that they meant that she herself acted more rationally at the time.

On my account of self-knowledge, Emily cannot pull out wholeheartedly because she does not know, at the time, what her real reason is for pulling out. She is confused. Her decision to pull out is conflicted. It is therefore psychologically unfree.

Huck Finn

Let us now also consider Arpaly's central case, Huck Finn, discussed by Jonathan Bennett (1974, pp. 123–134). Huck helps his friend Jim, a slave owned by Miss Watson, to run away. Huck, believing in the morality of the time, is certain that he should hand Jim in. His conscience makes him think, 'What had

poor Miss Watson done to you [to him, Huck], that you could see her nigger go off right under your eyes and never say a single word? What did that poor old woman do to you, that you could treat her so mean?'

He decides to hand Jim in and begins paddling to the shore to do so. Jim, exultant because he thinks he is being set free, tells Huck he will buy his children out of slavery or steal them. He says, 'Jim won't ever forgit you, Huck. You's de bes' fren' Jim's ever had…'.

And when two men hunting for runaway slaves ask Huck whether the man on his raft is black or white, he tries to say 'black' but 'the words wouldn't come'. He says 'white' instead, so protecting Jim. Bennet quotes Sidnell (pp. 205–206) who points out that for Huck, love and compassion for Jim are struggling against his conscience. But Bennett adds that 'to the end, Huck sees his compassion for Jim as weak, ignorant and wicked felony'.

Arpaly tells us that it is more plausible to say that Huck's decision not to turn Jim in

> is not only morally more admirable, but also more rational, than acting on his all-things-considered judgment. … By interfering with his better judgment the emotion enables Huck to track reasons he would have neglected had he followed the conclusion of his deliberation. (Tappolet 2007, p. 116)

I agree that it is admirable. It might also seem rational to us, in the twenty-first century, because in our culture we reject slavery as abhorrent. But Huck's action went against his own belief, the accepted belief of that time that he should turn Jim in. He let his emotion sway his judgment. This is an example of an alienated attitude (Huck sees his desire to help Jim escape as both weak and wicked[91]) interfering with his rational judgment. He struggles to do as his

[91] Bennett 1974, p. 126

conscience tells him and fails. He suffers ambivalence; his action in helping Jim escape, like Arpaly's case of Emily, is not fully self-determined and thus suffers a degree of passivity. We cannot *reflect on* unconscious self-knowledge, because we are necessarily unaware of it. Of course, its tacit, unconscious presence in our minds may affect our decision or our action causally.[92]

Conclusion

In this chapter I have continued with akrasia what I began in Chapter 5 with self-deception. In Chapter 5 I argue that inaccessible attitudes explain self-deception. In this chapter I argue that the anomaly of serious akrasia necessarily involves the subject's estrangement from her belief or intention on the ground that she necessarily cannot know *why* she believes that *p* or intends to do *a* in *all* of the following ways: consciously, immediately, wholeheartedly and in a settled way.

It follows that

- Her knowledge of this estranged attitude can be either conscious or unconscious.
- If it is unconscious, it is directly inaccessible to her; she is ignorant of it.

If it is explicitly conscious, then either

- the reason she has it is inaccessible to her (and so she is ignorant of this reason) or
- she knows this reason only theoretically—its phenomenological

[92] It is easy for us to use the accepted wisdom of our current society to condemn the actions of people who were brought up several decades earlier, under different understandings of certain activities that are (rightly) condemned today. I find this difficult to accept.

immediacy is inaccessible to her (and so she is ignorant of how this immediacy would feel).

In all of these situations, because of her estranged attitude, the subject is ignorant of why she does what she explicitly knows is against her better judgment. The attitude from which subjects are estranged can be accessible or inaccessible, conscious or unconscious. Either way, the subject is ignorant of some fact that is central to her action. Thus her self-ascription has only third-person authority.

This shows us that there is a big difference between self-deception and akrasia. When, in discussing akrasia, we use the concept of making up one's mind instead of the concept of willpower, we find that in every case the subject is conflicted between acting according to her better judgment and acting as she really wants to act.

In Section 1, I showed how Socrates, Aristotle, Davidson, Hare, Bratman and Mele attempt to answer the hard question: why would we ever act knowingly against our better judgment if we could resist doing so? I argue that in serious cases of akrasia, we never act *freely* against our better judgment; when we appear to do so, this is because although akratic subjects act as agents in such cases, their estranged attitude prevents their agency from being fully active. Their intention is thus never wholehearted, cannot be fully endorsed and has only third-person authority.

I conclude that weakness of will is not the best way of describing serious akratic cases. Everyday cases may or may not involve estrangement. Where they do not, and where the subject can fully accept her 'second-best' intention, she may not suffer conflict. Such cases might be better labelled non-akratic.

In Section 2 I defended my position on akrasia and used my Moran-based account of self-knowledge to consider an example that Tappolet has given us.

I argue that because she adopts a third-person, deflationary position towards her examples, Tappolet cannot use this example in the way she wants to—to show that emotions make akratic actions both intelligible and rational. My position is that any self-ascription that cannot be made wholeheartedly lacks first-person authority and active agency. The example of Emily shows us that, on my account, this subject's self-ascription has the authority only of the third-person perspective. Bratman's Sam, in Section 1, also has only third-person authority over his drinking.

Tappolet's discussion of the akratic case presented in this section is empirical and thus deflationary. Because empirical accounts of akrasia cannot distinguish between a third-person, theoretical stance and a first-person, avowable stance that a subject can adopt towards her conflicting attitudes, they cannot differentiate between self-ascriptions that have first-person agency and self-ascriptions that do not. Nor can they incorporate into their arguments the concept that unconscious, estranged, inaccessible attitudes do occur, do conflict with conscious attitudes, and do often cause akrasia, along with addictive and other compulsions and confusions. Only an account of self-knowledge from the first-person position of the active agent, such as Moran's, can differentiate between intentions that have first-person authority and intentions that have third-person authority only.

Moran's account of self-knowledge thus poses a problem for deflationary accounts of self-deception and akrasia. A number of writers now argue that willpower is a matter of being able to know one's own mind as an active agent in control of what one believes or intends to do, a position consistent with Moran's own approach to the anomaly of akrasia. The examples discussed in this chapter support these conclusions.

Thesis conclusion

The two main anomalies are self-deception and akrasia. I have set phobias aside because there was insufficient space in the thesis to accommodate the thorough examination that my explanation of them would require. My main aim has been to develop a general account of the anomalies by focusing on self-deception and akrasia.

My goal throughout has been to use Moran's account of self-knowledge to provide a more satisfactory account of the anomalies of self-deception and akrasia than common extant accounts. To do this, I first needed to lay out Moran's account and defend it against objections. In doing this, I presented my particular understanding of his account that I then applied to the anomalies. But first I needed to provide a defence of its main sub-accounts, those of rational agency, transparency and epistemic substantivity, against deflationary objections. Only having done all of that could I safely turn to the anomalies. In order to clarify the kinds of anomalous attitudes that occur, I have argued that we need to use all of Moran's major claims: his concepts of irreducibility, of estrangement and authority and of the relation between them, his distinction between stances, and his accounts of rational agency, transparency and substantivity, in order to distinguish between examples that carry first-person authority and examples that carry only third-person authority. Deflationist accounts of self-deception, for example, cannot allow the irreducible difference in perspectives between the subject's avowal 'I believe that *p*' and her tacit (unconscious) knowledge that she also, at the same time, believes that $\sim p$.

I have drawn the following conclusions concerning the anomalies.

Broadly, the concept of estrangement (i.e., of alienation, since I have used

'alienation' interchangeably with 'estrangement') gives us a better explanation, I have argued, of existing puzzles concerning self-deception and akrasia. Estrangement produces a degree of passivity in the subject that renders her self-ascription psychologically unfree because it cannot be fully self-determined; such self-ascriptions therefore merit only third-person authority. More specifically:

All the anomalies involve some examples where either i) the subject has an attitude from which she is estranged in the sense of being unable consciously to access it or ii) she has two conflicting attitudes, from both of which she feels consciously alienated to some degree. In self-deception, the subject always has the first kind of attitude. Serious akratic action also always involves this kind of estranged attitude. Everyday akratic acts are less likely to involve an attitude that is estranged in the sense of being inaccessible to the subject's consciousness; in such cases it is more likely that the akratic subject feels consciously alienated from both a and ~a because she feels conflicted about which of these to choose.

The self-ascription of a subject who has an estranged attitude in either of these ways necessarily involves a degree of passivity. This implies that it is not completely self-determined and thus is to some extent psychologically unfree. As such self-ascriptions are agential, they therefore lack the first-person authority that self-ascriptions of intentional attitudes normally have. In both kinds of example, the subject necessarily cannot know consciously, immediately, wholeheartedly and in a settled way why she believes that p or intends to do *a*.

Some self-ascriptions are said to be anomalous because they seem to be inexplicable; they are based on reasons that are neither justifiable nor explicable. I have argued that the anomalies are not anomalous at all; they are reasons- based; some aspect of one or more of these reasons, however, is not

available to the subject, who thus necessarily has only third-person authority over her self-ascription. Deflationary accounts of the anomalies, because they assume that the first-person perspective is reducible to the third, cannot provide us with a way of distinguishing between a subject's first-person authority over her belief that *a* and her intention to *a,* and her third-person authority over these. Subjects whose actions have only third-person authority often feel obliged to invent an alternative reason, either false or only accidentally true, for, say, their self-deceptive self-ascription. For example, Janet is estranged from her shame at having a daughter with learning difficulties; she reacts by accusing the daughter's teacher of treating her daughter unfairly in class. That way it is the teacher who is shameful, not Janet or her daughter. Janet is self-deceived. Her self-ascription is, however, still agential—it just suffers a degree of passivity on a Kantian account of active agency. Deflationary accounts of the anomalies cannot distinguish between the two stances of attribution and avowal, with their two different kinds of authority. I conclude that my interpretation provides a better and more natural reading of both the serious and the familiar cases I discuss than do either irrationality or other forms of ignorance.

References

Almeida, C 2001, 'What Moore's Paradox Is About', *Philosophy and Phenomenological Research*, vol. 62, no. 1, pp. 33 – 58.

Anscombe, G.E.M., 1957, *Intention*, 2nd edn (1976), Cornell University Press, Ithaca, NY.

Aristotle 1984, *Complete works of Aristotle*, 2 vols, ed. J Barnes, Princeton University Press, Princeton, NJ.

Aristotle 2011, *Nicomachean ethics*, trans. RC Bartlett & SD Collins, University of Chicago Press, Chicago, IL.

Arpaly, N 2000, 'On acting rationally against one's best judgement', *Ethics*, vol. 110, pp. 488–513.

Audi, R 1988, 'Self deception, rationalization and reasons for acting', in *Perspectives on self-deception*, eds BP McLaughlin & AO Rorty, University of California Press, Los Angeles, CA, pp. 92–120.

Bargh, JA, Chen M & Burrows L 1996, 'Automaticity of social behaviour: direct effects of trait construct and stereotype activation on action', *Journal of Personality and Social Psychology*, vol. 71, pp. 230–44.

Bar-On D & Long D 2003, *Privileged access: philosophical accounts of self-knowledge*, ed. B Gertler, Ashgate: Aldershot, UK.

Bayne, T & Fernandez, J 2009, 'Delusion and self-deception: mapping the terrain' in *Delusion and self-deception*, eds T Bayne & J Fernandez, Psychology Press, New York, pp. 1–22.

Bilgrami, A, 1998, 'Self-knowledge and resentment' in *Knowing our own minds*, eds C Wright, B Smith & C Macdonald, Oxford University Press, Oxford, pp. 207–242.

Boghossian, P, 2003, 'Content and self-knowledge' in *Privileged access: philosophical accounts of self-knowledge*, ed. Brie Gertler, Ashgate, Aldershot, UK.

Bortolotti, L & Broome, MR, 2008, 'Delusional beliefs and reason giving', *Philosophical Psychology*, vol. 21, no. 6, pp. 821–841.

Bortolotti, L 2010, *Delusions and other irrational beliefs*, Oxford University Press, Oxford.

Boyle, M 2009, 'Two kinds of self-knowledge', *Philosophy and Phenomenological Research,* vol. 78, no. 1, p. 133–164.

Boyle, M 2011, 'Transparent self-knowledge', *Proceedings of the Aristotelian Society,* vol. 85, pp. 223–241.

Bratman, M 1979, 'Practical reasoning and weakness of will', *Nous*, vol. 13, no. 2, pp. 153–171.

Burge, T, 1996, 'Our entitlement to self-knowledge', *Proceedings of the Aristotelian Society,* vol. 96, pp. 91–116.

Byrne, A, 2005, 'Introspection', *Philosophical Topics*, vol. 33, pp. 79–104.

Byrne, A, 2011, 'Transparency, belief, intention', *Proceedings of the Aristotelian Society*, vol. 85, pp. 201–21.

Canfield, J & Gustafson, D, 1962, 'Self-deception', *Analysis*, vol. 23, pp. 32–36.

Carman, T 2003, 'First-persons: on Richard Moran's *Authority and estrangement*', *Inquiry*, vol. 46, no. 3, p. 395–408.

Cassam, Q 2010, 'Judging, believing and thinking', *Philosophical Issues*, vol. 20, no. 1, pp. 80–95.

Child, W 2009, '*Authority and Estrangement: An Essay on Self-Knowledge*, by Richard Moran', *Mind*, vol. 118, no. 471, pp. 850-855.

Corbi, JE 2007, 'The mud of experience and kinds of awareness', *Theoria*, vol. 22, no. 1, pp. 5–15.

Corbi, JE 2009, 'First-person authority and self-knowledge as an achievement', *European Journal of Philosophy*, vol. 18, no. 3, pp. 325–362.

Davidson, D. 1984, 'First-person authority', *Dialectica*, vol. 38, no. 2–3.

Davidson, D 1980, 'How is weakness of will possible?' in *Essays on action and events*, ed. D Davidson, Clarendon Press, Oxford, pp. 21–42.

De Sousa, R 2002, 'Emotional truth', *Proceedings of the Aristotelian Society,* vol. 76, pp. 247–263.

Di Nucci, E 2012, 'Priming effects and free will', *International Journal of Philosophical Studies*, vol. 20, no. 5, pp. 725–734.

Dretske, F 1994, 'Introspection', *Proceedings of the Aristotelian Society,* vol. 94, pp. 263–278.

Edgeley, R 1969, *Reason in theory and practice*, Hutchinson, London.

Evans, G 1982, *The varieties of reference*, Oxford University Press, Oxford.

Falvey, K 2000, 'The basis of first-person authority', *Philosophical Topics*, vol. 28, no. 2, pp. 69–99.

Fernandez, J 2003, 'Privileged access naturalized', *The Philosophical Quarterly*, vol. 53, no. 212, pp. 352–372.

Fowler, HW & Fowler, FG (eds) 1964, *The concise Oxford dictionary*, 5th edn, Oxford University Press, London.

Friedrich, J 1993, 'Primary error detection and minimization PEDMIN strategies in social cognition: a reinterpretation of confirmation bias phenomena', *Psychological Review*, vol. 100, pp. 298–319.

Gallois, A 1996, *The world without, the mind within*. Cambridge University Press, Cambridge.

Gardner, Sebastian, 2004, 'Critical notice of Richard Moran, *Authority and estrangement: an essay on self-knowledge*', *The Philosophical Review*, vol. 113, No. 2.

Gertler, B 2001, Introspecting phenomenal states, *Philosophy and Phenomenological Research*, vol. 63, pp. 305–328.

Gertler, B 2009, entry, 'Self-knowledge', *Encyclopedia of philosophy*, ed. EN Zalta.

Gertler, B 2011, 'Self-knowledge and the transparency of belief', in *Self-knowledge*, ed. A Hatzimoysis, Oxford University Press, Oxford, pp. 125–145.

Gordon, R 1986, 'Folk psychology as simulation', *Mind and Language*, vol. 1, pp. 158–71.

Green M. & N. Williams John (eds), 2007, *Moore's Paradox: New Essays on Belief, Rationality, and the First Person*, Oxford University Press, Oxford.

Hare, RM 1952, *The language of morals*, Clarendon Press, Oxford.

Hare, RM 1992, entry, 'Weakness of will', in *Encyclopedia of ethics*, ed. L Baker, Garland, NY, pp. 1304–1307.

Jackson, F 1984, 'Weakness of will', *Mind*, vol. 93, pp. 1–18.

Kant, I 1997, *Critique of pure reason*, ed. & trans. P Guyer & AW Wood, Cambridge University Press, Cambridge.

Macdonald, C 1998, 'Externalism and authoritative self-knowledge', *Knowing our own minds*, eds C Wright, BC Smith & C Macdonald, Oxford University Press, Oxford, pp. 123–154.

Marshall, G 2000, 'How far down does the will go?' in *The analytic Freud*, ed. MP Levine, Routledge, London, pp. 36–48.

Martin, MGF 1998, 'An eye directed outward', in *Knowing our own minds*, eds C Wright, BC Smith & C Macdonald, Oxford University Press, Oxford, pp. 64–98.

McDowell, J, 1998, 'Response to Crispin Wright', in *Knowing our own minds*, eds C Wright, BC Smith & C Macdonald, Oxford University Press, Oxford, pp. 14–45.

McLaughlin, B, 1988, 'Exploring the possibility of self-deception in belief', in *Perspectives on self-deception*, eds BP McLaughlin & AOK Rorty, University of California Press, Los Angeles, CA, pp. 29–62.

Mele, AR 1987, *Irrationality: an essay on akrasia, self-deception and self-control,* Oxford University Press, New York, NY.

Mele, AR 1991, 'Akratic action and the practical role of better judgment', *Pacific Philosophical Quarterly*, vol. 72, pp. 33–47.

Mele, AR 1997, 'Real self-deception', *Behavioral and Brain Sciences*, vol. 20, pp. 91–136.

Mele, AR 1999, entry, *The Cambridge dictionary of philosophy*, ed. R Audi, 2nd

edn, p. 16.

Mele, AR 2002, 'Akratics and addicts', *American Philosophical Quarterly*, vol. 39, no. 2, pp. 153–167.

Mele, AR 2009, 'Self-deception and delusions', in *Delusion and self-deception*, eds T Bayne & J Fernandez, Psychology Press, New York, NY, pp. 55–69.

Moran, R 2001, *Authority and estrangement: an essay on self-knowledge*, Princeton University Press, Princeton, NJ.

Moran, R, 2003, 'Responses to O'Brien and Shoemaker', *European Journal of Philosophy*, vol. 11, no. 3, pp. 402–419.

Moran, R, 2004a, 'Precis of *Authority and estrangement: an essay on self–knowledge*', *Philosophy and Phenomenological Research*, vol. 69, no. 2, pp. 423–426.

Moran, R, 2004b. 'Replies to Heal, Reginster, Wilson, and Lear', *Philosophy and Phenomenological Research*, vol. 69, no. 2, pp. 455–472.

Moran, R, 2007, 'Replies to critics', *Theoria*, vol. 58, pp. 53–77.

Moran, R, 2012, 'Self-knowledge, "transparency", and the forms of activity', in *Introspection and consciousness*, eds D Smithies & D Stoljar, Oxford University Press, New York, NY.

O'Brien, L 2003, 'Moran on self-knowledge', *European Journal of Philosophy*, vol. 11, no. 3, pp. 391–401.

Owens, D 2000, *Reason without freedom*, Routledge, London.

Owens, D 2003, 'Knowing your own mind', *Dialogue*, vol. 42, pp. 971–998.

Paul, S 2012, 'How we know what we intend', *Philosophical Studies*, vol. 161,

no. 2, pp. 327–346.

Peacocke, C 1998, 'Conscious attitudes and self-knowledge', in *Knowing our own minds*, eds C Wright, BC Smith & C Macdonald, Oxford University Press, Oxford, pp. 64–98.

Peacocke, C 1992, *A study of concepts*, Harvard University Press, Cambridge, MA.

Petocz, Agnes 1999, *Freud, psychoanalysis and symbolism*, Cambridge University Press, Cambridge.

Pink, T 1996, *The psychology of freedom*, Cambridge University Press, Cambridge.

Prades, J 2007, 'Endorsement, reasons and intentional action', *Theoria*, vol. 58, pp. 25–33.

Reed, B, 2010, 'Self-knowledge and rationality', *Philosophy and Phenomenological Research*, vol. 80, no. 1, pp. 164–181.

Reginster, B 2004, 'Self-knowledge, responsibility and the third person', *Philosophy and Phenomenological Research*, vol. 69, no. 2, pp. 433–439.

Rey, G 1988, 'Towards a computational account of akrasia and self-deception', in *Perspectives on self-deception*, eds AO Rorty & B McLaughlin, University of California Press, Los Angeles, CA, pp. 264–296.

Sanford, D 1988, 'Self-deception as rationalization', in *Perspectives on self-deception*, eds B McLaughlin & A O Rorty, University of California Press, Los Angeles, CA, pp. 157–169.

Sellars, W 1997, *Empiricism and the philosophy of mind*, Harvard University Press, Cambridge, MA.

Shah, N & Velleman, D 2005, 'Doxastic deliberation', *The Philosophical Review*, vol. 114, no, 4, pp. 497–534.

Shoemaker, S 1994, 'Self-knowledge and inner sense', *Philosophy and Phenomenological Research*, vol. 54, pp. 249–314.

Shoemaker, S 2003, 'Moran on self-knowledge', *European Journal of Philosophy*, vol. 11, no. 3, pp. 391–401.

Shoemaker, S 2003, 'Special access lies down with theory-theory', *Behavioral and Brain Sciences*, vol. 16, pp. 77–79.

Shoemaker, S 2009, 'Self-intimation and second order belief', *Erkenntnis*, vol. 71, pp. 35–51.

Sims, A 2003, *Symptoms in the mind*, Elsevier, London.

Stanley, J 2011, *Know how*, Oxford University Press, Oxford.

Stroud, S 2009, entry, 'Weakness of will', *Stanford Encyclopedia of philosophy,* ed. EN Zalta, Stanford University Press, Stanford, CA, p. 1.

Stroud, S & Tappolet, C 2007, 'Introduction', in *Weakness of will and practical irrationality*, eds S Stroud & C Tappolet, Oxford University Press, New York, NY, pp. 1–16.

Stroud S & Tappolet, C (eds) 2007, *Weakness of will and practical irrationality*, Oxford University Press, New York, NY.

Tappolet, C 2007, 'Emotions and the intelligibility of akratic action', in *Weakness of will and practical irrationality,* eds S Stroud & C Tappolet, Oxford University Press, New York, NY, pp. 97–120.

Taylor, C 1985, 'The concept of a person', in *Human agency and language: philosophical papers*, Cambridge University Press, Cambridge, vol. 1, pp.

97–114.

Tenenbaum, S 1999, 'The judgment of a weak will', *The Philosophical Review*, vol. 86, pp. 316–339.

Velleman, D 2000, *The possibility of practical reason*, Oxford University Press, Oxford.

Wallace, RJ 1990, 'How to argue about a practical reason', *Mind*, vol. 99, pp. 355–385.

Watson, G 1977, 'Skepticism about weakness of will', *Philosophical Review*, vol. 86, pp. 316–313.

Watson, G 2007, 'The work of the will', in *Weakness of will and practical irrationality,* eds S Stroud & C Tappolet, Oxford University Press, New York, NY, pp. 172–200.

Way, J 2007, 'Self–knowledge and the limits of transparency', *Analysis*, vol. 67, no. 3, pp. 223–230.

Wittgenstein, L 1974, *Philosophical investigations*, trans GEM Anscombe, Blackwell, Oxford.

Wittgenstein, L 1980*, The blue and the brown books*, Blackwell, Oxford.

Wright, C 1989, 'Wittgenstein's later philosophy of mind: sensation, privacy and intention', *Journal of Philosophy*, vol. 86, no. 11, pp. 622–634.

Wright, C 1996, 'On making up one's mind: Wittgenstein on intention', in *Logic, philosophy of science and epistemology:* proceedings of the 11th International Wittgensteinian Symposium, eds P Weingartner & G Schurz, Kirchberg, Vienna.

Yager, J & Gitlin, MJ 2005, 'Clinical manifestations of psychiatric disorders', in

Kaplan and Sadock's comprehensive textbook of psychiatry, 8th edn, eds BJ Sadock & VA Sadock, Lippincott, Williams & Wilkins, Philadelphia, PA, vol. 1, pp. 1964–1002.

Zimmerman, A 2004, 'Unnatural access', *The Philosophical Quarterly*, vol. 54, no. 216, pp. 435–438.