2010

Self-Knowledge and Rationality

Stephen Blackwood
Wilfrid Laurier University

Your file Votre référence
ISBN: 978-0-494-64402-7
Our file Notre référence
ISBN: 978-0-494-64402-7

SELF-KNOWLEDGE AND RATIONALITY

By

Stephen Blackwood

M.A., McMaster University, 2002

B.A. (Hon.), Memorial University of Newfoundland, 1991

DISSERTATION

Submitted to the Department of Philosophy

in partial fulfillment of the requirements for

Doctor of Philosophy

Wilfrid Laurier University

2010

Abstract

Several basic asymmetries are normally thought to exist between first- and third-person present-tense ascriptions of mental states. First of all, when a speaker ascribes, for instance, a belief that *p* to another, she must do so on the evidence provided by the utterances and actions of the other. However, it at least appears that typically she need not do so when ascribing a belief to herself. In other words, there is an *immediacy* to a self-ascription of a belief (that is, an utterance of the form 'I believe that *p*') that third-person ascriptions ('He believes that *p*') lack. Secondly, our self-ascriptions are *groundless* - demands that we justify our self-ascriptions, or explain how we know that we are in the mental states we self-ascribe, are generally deemed inappropriate. Thirdly, assuming sincerity on the part of the speaker, a self-ascription of a mental state is highly likely to be correct. This likelihood of correctness is not thought to extend to her ascription of similar beliefs to others. Thus, it is claimed, speakers possess a level of *authority* with respect to their self-ascriptions that they do not enjoy with regard to their attribution of beliefs to others.

Discussions of 'the problem' of self-knowledge often focus on these asymmetries and the prima facie tension between the idea that the first person needs none of the evidence on which the third person depends, and yet is more likely to be correct. In what does this apparently special way of knowing our own minds consist? In recent times a number of philosophers (for example, Sydney Shoemaker, Tyler Burge, Akeel Bilgrami, Richard Moran and Dorit Bar-On) have pursued this goal by linking self-knowledge claims (authoritative self-ascriptions of mental states) to the critical rationality and rational

agency taken as essential to the first-person perspective. While their approaches differ in various respects, each argues that (1) self-ascriptions express second-order beliefs about first-order mental states, and that (2) the explanation of the truth of, and warrant for, these beliefs that qualifies them as knowledge is to be found in the requirement for self-knowledge that the possibility of rationality demands.

Looking at how (1) is understood is essential for assessing the plausibility of this normative turn in the explanation of self-knowledge, and arguments for a substantial epistemic account of self-knowledge more generally. Determining in what sense, if any, (i) self-ascriptions may be thought to count as expressions of second-order beliefs, and (ii) the role second-order belief might play in securing the truth of self-ascriptions, will have consequences for understanding what role, if any, normative second-order judgement (that is, judgement about what first-order state one ought to have) may play in what is normally called self-knowledge. I argue that various problems with the views of each of the philosophers mentioned above points to the need for a non-epistemic explanation of our authoritative self-ascriptions, where such self-ascriptions are taken as expressive not of second-order beliefs about our mental states, but of the first-order states they semantically specify. I contend that a good account can be found through combining Davidson's explanation of first-person authority with an expressivist reading of the first-order expressive character of self-ascriptions.

With an epistemically deflationary explanation of authoritative self-ascription in place, what becomes of the understanding of rationality argued for by Shoemaker et al.? Following David Owens, I first argue that, even if we were to possess the kind of self-

knowledge these philosophers suppose us to have, we could not exercise the kind of higher-order control over our first-order states for which they argue. I then close out the discussion by offering an outline of an alternative conception of rationality – that of Donald Davidson – that points to how we may conceive of rationality without self-knowledge.

Acknowledgements

To Marie-Claude, for her unending love and support. Also to Anna and Sam, who came along along the way, and fill every day with surprise and joy. Also to my parents, Ches and Ann, for all their love and encouragement o'er lo these many years. My thanks to the Examination Committee members: Dr. Mark McCullagh, Dr. Jill Rusin, Dr. Michael Hymers, Dr. Lauren King, and Dr. Tobias Krettenauer. Finally, deepest thanks to my advisor, Dr. Rockney Jacobsen, who went far beyond the call of duty in ways too numerous to mention.

Table of Contents

Chapter 1: The Problem of Self-Knowledge

Introduction

In this chapter I offer an overview of some of the major issues that have been featured in recent discussions of what has come to be called the problem of self-knowledge. I begin with a description of some of its distinguishing features, such as the asymmetries that are thought to obtain between self- and other-ascriptions of mental states. I then look at Paul Boghossian's instructive discussion of some of the various approaches that might be taken in accounting for these features. Finally, I move to an introductory discussion of a number of different approaches that philosophers have taken in their attempts to come to grips with them. In particular, I focus on a relatively recent trend among some philosophers to account for self-knowledge by explaining the link it bears to rational agency. I also introduce an alternative non-epistemic account – one that argues for an understanding of self-ascriptions as expressive not of second-order beliefs about our mental states, but of the first-order states they semantically specify – that undermines this connection. This will introduce the central characters and ideas that will come in for criticism in Chapters 2 and 3, as well as the positive views that will begin to emerge in Chapter 3 and will be defended in Chapter 4.

1.1 Some Basic Asymmetries Between Self- and Other-Ascriptions

Our first-person present-tense ascriptions of contentful mental states (for example, of belief, desire, intentions), and phenomenal states (such as pains and the like) are thought to differ in a number of significant and fundamental ways from our ascriptions of those states to others. For example, when a person ascribes a mental state that *p* to another, she must do so on the evidence provided by the utterances and actions of the other. However, it at least appears that typically she need not do so when ascribing such states to herself. In other words, there is what we will call an *immediacy* to a first-person present-tense ascription of, for example, a belief (that is, of the form 'I believe that *p*') that third-person and past-tense ascriptions ('He believes that *p*') lack. In addition, unlike other-ascriptions, self-ascriptions are typically taken to be *groundless*, in the sense that demands that we justify our self-ascriptions, or explain how we know that we are in the mental states we self-ascribe, are generally deemed inappropriate. Furthermore, assuming sincerity on the part of the person, such self-ascriptions (those not ascribed on the basis of behavioural evidence) are highly likely to be correct. This likelihood of correctness is not thought to extend to her ascriptions of similar mental states to others, or to past-tense ascriptions to ourselves. Thus, it is said, persons appear to possess a level of *authority* with respect to certain of their self-ascriptions that, while it falls short of infallibility, is far greater than that which they enjoy with regard to their attribution to others.

Amongst those who accept that these asymmetries obtain, two general explanatory paths have been taken. In more recent times, an *epistemically deflationary* approach has

gained some currency. According to this view, the authority and immediacy generally granted to certain kinds of self-ascriptions are not to be explained in terms of any privileged position the subject occupies with respect to the perception of her mental states, nor in any advantage she might enjoy with respect to the amount or quality of evidence she might have for them. Instead, it is based on some other non-epistemic feature of self-ascriptions. Still, for most philosophers the question remains an epistemic one. The task as these philosophers see it is to show how we may incorporate the asymmetries into an account that explains how self-ascriptions express knowledge, that is, as a form of justified true second-order belief about first-order mental states.

1.2 An Epistemically Oriented Analysis of the Problem of Self-Knowledge

An example of an epistemically oriented examination of this issue is Paul Boghossian's essay 'Content and Self-Knowledge' (1998). I begin with a look at this essay because it serves as a good starting point for getting a sense of what sort of phenomena and problems are often associated with the subject of self-knowledge. Furthermore, through an analysis of his claims we can begin to develop an idea of what direction a resolution of these problems might take.

According to Boghossian, the basic issue is this. On the one hand, the idea of self-knowledge – the capacity to formulate justified true beliefs about our mental states – is presupposed by many of the concepts (for example, intentional action) that are fundamental to our ordinary self-conception. Consequently, insofar as we cannot see our

way to an alternative self-conception, a skeptical view that denies such a capacity must be rejected. On the other hand, upon inspection we find that each of the various options for an epistemic account comes up wanting. The conclusion is that, while we cannot do without the idea of self-knowledge, as of yet we have little idea what form an epistemic explanation consistent with the characteristic features and related epistemic norms associated with self-knowledge judgements might take.

Boghossian arrives at this conclusion after examining epistemic problems generated by apparently irreconcilable features of self-knowledge. The general question Boghossian addresses is how to account for our capacity to produce true justified beliefs about our thoughts, where that includes not just the thought that *p,* but mental states such as the belief, desire, or fear that *p* as well. We can begin by looking at an example of "everyday" self-knowledge that Boghossian offers early on. He writes that, immediately upon thinking 'Even great composers write lousy arias,' one knows what one has thought (Boghossian, 152). Presumably this means something like: One has immediate understanding of what one is thinking that one could manifest in a true justified second-order judgement self-ascribing the content and type of mental state in question (for example, 'I believe that even great composers write lousy arias'). As Boghossian sees it there are three possible avenues an explanation of our capacity for self-knowledge might follow. One could show how such judgements are derived from (1) inference of some kind, (2) some sort of inner observation, or (3) some other non-empirical basis (ibid., 149-150).

If we look at the inferential option, we see that, for starters, it would seem to go against our epistemic intuitions regarding the immediacy of self-ascriptions outlined above. Beyond that, for many self-ascriptions the type of behavioural evidence to which an inferential account would have to appeal is not available to the thinker at the time the thought is made. For example, sitting quietly at my desk I might think 'Even great philosophers sometimes make mistakes,' immediately upon which, according to Boghossian I would know that I thought so in spite of lacking behavioural evidence that might manifest the thought and serve as premises for an inference to its self-ascription. In addition to this, Boghossian contends that an internalist conception of justification, to which many philosophers remain sympathetic, demands that self-knowledge be non-inferential. On the internalist view of justification, one may be justified in one's belief that one believes that *p* only if one recognizes (i) the belief upon which that belief rests (a belief that q), as well as (ii) that one believes it. He outlines it as follows:

(1) I believe that I believe that r.

(2) I believe that s.

(3) The proposition that s justifies the proposition that I believe that r.

(4) I know that I believe that s.

(5) I know that a belief that s justifies that I believe that r.

(6) I believe that I believe that r as a result of the knowledge expressed in 4 and 5.

(ibid., 155)

The difficulty is that on the internalist view, the justification of (1) requires that I already know that I have certain beliefs, as is evident in condition (4). But then there is

the question of the justification of those beliefs (and so on), which sends us off on a vicious regress. We are left to conclude that there must be some way to know the content of one's mental states (including thoughts) non-inferentially. This leaves us with the remaining two possibilities: either self-knowledge is based on inner observation, or it is grounded on "nothing empirical" (ibid., 156).

The inner observation option, while perhaps not so immediately counter-intuitive, is also untenable. The idea here is that, given certain widely accepted externalist claims about the character of thought content, it follows that we could not know the content of our thoughts through mere inspection of their intrinsic (narrow) properties. To know that one is thinking of water, and not *twater*, one needs to know its relational property, for example, that one's thought is caused by H_2O and not A_2Z. However, no inner observation or introspection of the intrinsic properties of that thought will give one the requisite knowledge of that extrinsic property. Consequently any judgement about what we are thinking will be susceptible to the skeptical charge that we don't know what content we are attributing to ourselves – that is, we lack semantic authority with respect to the meanings of the terms through which we express our thoughts (ibid., 166). So, in brief, the argument goes. But if this is correct, then we are left with our third option, that self-knowledge is based on nothing. What does he mean by this?

Normally our knowledge of a contingent proposition is grounded on observation or some inference based on some observation. As Boghossian puts it, such empirical knowledge involves a "cognitive achievement," and its epistemology is always "substantial" (Boghossian, 165). Knowledge that is "based on nothing" does not derive

from any such cognitive achievement and its epistemology is therefore "insubstantial." Boghossian offers a few examples of potentially baseless, or insubstantial, knowledge. First, there are certain self-regarding indexical propositions, such as "I am here now," that are true and justified as soon as thought. Secondly, some philosophers argue that in some cases there are self-regarding, self-verifying propositions that, on being thought, constitute one as being in the state they indicate. For example, there may be no fact of the matter about my being jealous of my friend prior to my judgement that I am, but my sincerely thinking it makes it so. In these cases, such judgements would be both true and justified, even though they were not grounded on any empirical evidence – observation, or inference from observation, would be irrelevant to the question of their truth or warrant.[1]

A third sort of insubstantial self-knowledge claim Boghossian considers is Tyler Burge's "basic self-knowledge" – self-ascriptions of the form 'I am thinking that *p*'.[2] Burge argues that in thinking such a second-order judgement one also thinks the first order judgement (that *p*) that it is about. He thinks this overcomes the problem of our authority regarding our knowledge of thought content outlined above – because of their self-referential, logically self-verifying character, one need not have "absolute" authority with respect to thought content for such judgements to count as instances of (insubstantial) self-knowledge. Boghossian does not disagree; however, he observes that

[1] This type of account bears similarities to Crispin Wright's constitutive account (which will be discussed in Chapter 3); however, Wright denies the need for any kind of explanation of epistemic warrant precisely because he thinks such judgements do not involve cognitive achievement.

[2] See Burge: 1998c.

such an analysis does not apply to judgements concerning a variety of propositional attitudes, because one need not actually believe, desire, or fear that *p* to think (make the judgement) 'I believe/desire/fear that *p*' (ibid., 169).[3]

The limited scope of each of these accounts points to the difficulty one faces in arriving at an insubstantial explanation of the general authority we are said to enjoy with regard to our thoughts. However, for Boghossian this lack of general application is not the most pressing issue such accounts face. As he sees it, the main problem is that the truth of judgements of the kinds mentioned is guaranteed. But this, he argues, is not in keeping with our ordinary conception of self-knowledge – authority is not thought to equal infallibility. He writes: "I know of no convincing alternative to the following type of explanation: 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 relevant evidence – and not" (ibid., 167). It would appear, then, that we must make room for "genuine cognitive achievement" in our account of self-knowledge after all, for otherwise we will have no way of making sense of our admitted imperfection in this regard.

It seems that we are in a quandary – we are left to conclude that while our ability to make knowledgeable judgements about our mental lives must involve cognitive achievement, all of the possibilities considered fall short. Again, this is not to say that Boghossian thinks a solution is impossible – he is optimistic that some version of one of

[3] It should be mentioned that Burge has recognised the limited application of his analysis and has subsequently offered a quite different sort of explanation – to be discussed below – of the knowledgeable status of judgements of the sort Boghossian mentions.

the options will work. Still, it remains that "we have a serious problem explaining our ability to know our own thoughts, a problem that has perhaps not been sufficiently appreciated" (ibid., 172).

To recap, according to Boghossian a theory of self-knowledge must include an account of:

(1) the immediacy of self-ascriptions;

(2) semantic authority;

(3) *how* we successfully self-ascribe mental states;

(4) the highly secure yet fallible character of self-ascriptions;

(5) the grounds on which true self-ascriptive second-order beliefs are justified.

The last in this list is tied to his initial anti-skeptical claim that our ordinary self-conception, which we cannot as of yet conceive of doing without, presupposes the ability to make knowledgeable judgements about our own mental states, together with the assumption that knowledge is true justified belief. In fairly recent times the idea that self-knowledge is essential to our self-conception has become the focus of a number of philosophers' attempts to explain the asymmetrical character of self- and other-ascriptions of mental states. Sydney Shoemaker, Tyler Burge, Akeel Bilgrami, and Richard Moran have each argued for an essential link between the authority that is thought to accrue to self-ascriptions and our status as rational subjects. In his own way each argues that an understanding of how our self-ascriptions count as knowledge is to be found in consideration of the role first-person second-order judgement and belief plays in

rational agency. Since these are the philosophers whose ideas will be central to the discussion in the chapters to come, I shall now give a brief overview of each view.

1.3 Shoemaker: The Necessity of Self-Awareness for Rationality

In 'On Knowing One's Mind' (1996a), Shoemaker contends that the rationalisation of the modification of belief requires self-knowledge ("or at least something very much like it," as he puts it [Shoemaker: 1996a, 31]). More specifically, it requires (1) second-order beliefs about what one's current first-order beliefs and desires are, (2) second-order desires to promote consistency in those first-order beliefs, and (3) second-order beliefs regarding what changes would be required in order to satisfy those second-order desires (ibid., 33). Furthermore, he offers a *reductio* argument against a phenomenon that he calls 'self-blindness' (a condition wherein one could recognize the truth of one's second-order beliefs only through interpreting one's own behaviour) to show that the kind of knowledge of one's first-order mental states needed must be gained via a kind of immediate privileged access he terms 'self-acquaintance' (ibid., 25). The argument goes like this: If self-knowledge by self-acquaintance were an optional component of our rational lives – in other words, if self-blindness were possible – then in cases in which a self-blind person lacked self-knowledge that could be gained only by self-acquaintance it would reveal itself in discrepancies between her behaviour and the behaviour of one who possessed such knowledge (a normal person, as he puts it). However, he argues, upon investigating the possibility it turns out that no such discrepancy would be found. This

leaves us with two options: (1) deny that we actually do have self-knowledge by self-acquaintance, or (2) given the apparent absurdity of such a thought, take the fact that no difference could be discerned as a *reductio* of the possibility of self-blindness and thus proof of the necessity of privileged self-knowledge (ibid., 36, 39).

In effect, Shoemaker argues for the necessity of special second-order judgements about our beliefs and desires from the requirements for interpreting one another as engaging in rational deliberation. But there is more to the story – as will be discussed in Chapter 4, Shoemaker suggests that the kind of second-order judgement required for the modification of one's mental states is also the mechanism through which we express our agency. We are responsible for our beliefs and other mental states in virtue of the fact that we can exercise control over them through our second-order deliberations on their rational standing (ibid., 28). Given that this requires knowledge of what those states are, it follows that self-knowledge is essential to our status as rational agents. In what follows I shall refer to accounts of self-knowledge that link together the monitoring or regulative role of second-order belief and agency as *supervisory models of self-knowledge.* As I read him, Tyler Burge also subscribes to this sort of view.

1.4 Burge: Self-Knowledge and the Requirements of Critical Rationality

In 'Our Entitlement to Self-Knowledge' (1998c) Burge also takes second-order belief expressive of self-knowledge to be a fundamental component of critical rationality. He argues that the truth and warrant of second-order judgements constitutive of self-

knowledge is connected to the entitlement we have to knowledge claims in general. This is because critical reason is an essential component of the knowledge enterprise. That said, he also argues that the kind of entitlement attached to second-order judgements must be distinct from that of ordinary perceptual belief. As he puts it, "there must be a non-contingent, rational relation between relevant first-person judgments and their subject matter or truth," a relation that is constitutive of critical reason (Burge: 1998c, 246). More specifically, our entitlement to self-knowledge claims is tied to our status as critical reasoners, to our ability to operate in accord with norms of reason, even if these norms cannot be articulated by the reasoner him- or herself.

With respect to our reflective second-order beliefs in particular, our entitlement to them derives from the role they play in critical reason, from the fact that they add an essential element to the reasonability of the whole process of critical reasoning. If our judgements about our first-order mental states and their interrelations were not rational (that is, if we lacked entitlement to them), then our reflection on those states would fail to add to the rationality of the whole reasoning process. But, Burge says, "reflection does add a rational element to the reasonability of reasoning. It gives one rational control over one's reasoning." As he goes on to say, "critical reasoning just is reasoning in which norms of reason apply to how attitudes should be affected partly on the basis of reasoning that derives from judgments about one's attitudes" (ibid., 249). Thus, our status as critical reasoners confers epistemic entitlement on our second-order judgements about our first-order beliefs. However, Burge adds, entitlement is not enough – for similar reasons those second-order judgements must also be generally true; otherwise the link between the two

levels of belief, and consequently one's ability to reflect critically, would break down. If reflection bore on the truth of our second-order beliefs in a merely contingent way, then the reason-guiding and coherence-making functions of critical reflection would fail. Or if we were entitled to our second-order judgements but they were systematically mistaken, then we could not be critical reasoners. "For critical reasoning requires rational integration of one's higher-order evaluations with one's first-order, object-oriented reasoning. ... If the two came radically apart, or were only accidentally connected, critical reasoning would not occur" (ibid., 250).

So for Burge, self-ascribing a mental state knowledgeably is a basic component of critical reflection; if self-ascriptive judgements weren't reliably correct, then the critical reflection in which we engage could not get off the ground. Furthermore, like Shoemaker, Burge sees this second-order capacity as essential to agency – we can be held responsible for our mental states only because we are capable of reviewing our reasons and reasoning (ibid., 258).

1.5 Bilgrami: Our Concept of Self-Knowledge is Deeply Normative

Although he approaches the problem from a somewhat different angle, Akeel Bilgrami (1999) also sees a strong connection between self-knowledge and rational agency. According to Bilgrami, considerations of agency, which, following Strawson, he takes to be a thoroughly normative idea, conceptually account for self-knowledge. This, he argues, makes the very idea of self-knowledge a thoroughly normative concept. As he puts it, "there is no understanding [agency] in strictly metaphysical and non-normative

terms, as traditional discussions have assumed" (Bilgrami, 214). In defense of this claim he argues that the first-order mental states picked out by second-order reports are just those states that lead to conclusions or actions that can be the object of the 'internally justifiable reactive attitudes' characteristic of critical reason and rational agency (ibid., 219). So, like Shoemaker and Burge, Bilgrami argues that it is the role that our second-order states play in critical rationality, which is in turn partly constitutive of our notion of agency, that serves as the warrant for those higher-order judgements. Self-knowledge is a necessary condition for the implementation of practices surrounding assignments of responsibility and the reactive attitudes they express, for it is only when self-knowledge is present that assignments of punishment, blame, or praise are deemed appropriate.

With Bilgrami we see the epistemic emphasis shift even more explicitly toward our responsibility for and control of our first-order mental states exercised through our higher-order judgements about them. It is part and parcel of our self-conception as rational agents that we think of ourselves (and others) as having self-knowledge. This is also why he thinks a causal/perceptual account of self-knowledge will not suffice, because such accounts must allow for the possibility of a breakdown in the contingent relation between first- and second-order beliefs that our understanding of ourselves as agents cannot in principle accommodate (ibid., 210). The concepts of agency and responsibility, and not just rationality (which, on its own, might be conceived "mechanistically"), need to be placed at the center of an account of the special character of self-knowledge, for it is the "*activity* of, the *agency* involved in making certain kinds of rational judgements that presupposes self-knowledge" (ibid., 237). For a person to

have thoughts properly conceived requires that he have higher-order reactive attitudes towards them. A wholly passive 'thinker', one who only had thoughts assail him (which Bilgrami equates with the causal view), would not be a genuine thinker.

One of the distinguishing (and more controversial) features of Bilgrami's discussion is the claim that when a certain condition is met, individuals are infallible with respect to their avowals. When a self-ascription is made under 'the condition of responsible agency' – where an individual's conclusions or actions derive from theoretical or practical deliberation that can be the object of internally justified reactive attitudes – then it follows that "for each such state, its possessor believes that she has it, and has it if she believes that she has it" (ibid., 226). The implications and plausibility of this thesis will be addressed in Chapter 2.

1.6 Moran: The Importance of a Non-Alienated First-Person Perspective

Shoemaker, Burge, and Bilgrami share the assumption that our authoritative self-ascriptions express self-knowledge, and thus justified, true, second-order beliefs. In varying degrees and ways, each argues that the distinctive character of the self-ascriptions taken as expressive of these second-order beliefs – their immediacy, groundlessness, and unparalleled security – as well as their justification or warrant, are to be explained in terms of inherent links between self-knowledge and rational agency, the latter of which is construed in terms of the rational control the subject exercises over his mental life through his second-order deliberation on it. Thus, in Chapter 2 I group these philosophers together and offer a more detailed examination of their respective views.

Like these philosophers, Richard Moran (2001) links a proper understanding of the nature of self-knowledge to our capacity for rational deliberation and agency. He is sympathetic to the general tenor of their views, arguing that a proper discussion must go beyond an explanation of the special mode of awareness and security characteristic of avowals: "[t]he special features of first-person awareness cannot be understood by thinking of it purely in terms of epistemic access. ... Rather we must think of it in terms of the special responsibilities the person has in virtue of the mental life in question being his own" (Moran, 32). However, he argues that the scope of the explanations they offer is too restricted, because as they fail to fully account for the nature of what he calls "genuine" first-person awareness and knowledge of one's own beliefs.

According to Moran, authentic self-knowledge requires that one see one's beliefs and other attitudes as "expressive of his various and evolving relations to his environment, and not as a mere succession of representations (to which, for some reason, he is the only witness" (ibid., 32). As he sees it, talk of consciousness ought to entail more than a description of how we know our own minds. It carries with it a whole host of other implications for the subject and her responsibilities and commitments – the epistemic perspective she takes toward herself has significant consequences for her relation to herself and her self-conception. Consequently, a key issue is deliberation and the role it plays in self-constitution, in *making up* one's mind about what one ought to and will believe.

Moran's argument for a substantial epistemology focuses on what he understands to be the significant cognitive achievement involved in instances of genuine self-

knowledge. He thinks this requires that we broaden our conception of what the relevant asymmetries are between the first- and third-person perspectives. In effect, he seeks to widen the sense in which first-person authority should be considered – it should concern not merely our ability to get our mental states right, but the kind of control we should exercise over our mental states (ibid., 3-4). This change of focus marks a departure from the kind of argument offered by Shoemaker, Burge, and Bilgrami. Like them, Moran sees self-knowledge as intimately connected to rational agency. However, for him the key issue is what sort of *commitment* our self-ascriptions must express if they are to be genuinely authoritative. Instead of focusing on the role self-knowledge plays in the supervision of one's first-order states (and the control exercised therein), Moran argues that the defining mark of genuinely authoritative self-ascription is that it expresses a commitment to the state self-ascribed being determined by the subject's understanding of the first-order reasons for it. And it is in virtue of this *commitment* that she exercises control over her mental life and counts as an agent (ibid., 148-151) .

As Moran sees it, this "first-personal" aspect of self-knowledge has been completely left out of most previous discussions. Most accounts of self-knowledge have described it as something that could just as well be an ordinary third-person phenomenon, imported into a closed mental interior (for example, the internal theatre of Descartes, Locke, and Hume). Even those who have argued against the very possibility of self-knowledge have done so under the tacit assumption that, if there is to be such a thing, it must conform to a Cartesian-inspired model of introspection. In opposition to both these views, Moran seeks to develop an explanation of first-person awareness of one's mental life which is

"substantial, representing a genuine cognitive achievement, but which nonetheless decisively breaks with the Cartesian and empiricist legacy" (ibid., 3).

Contrary to epistemically deflationary views, which he interprets as arguing against the idea of a fully independent object that could serve as the object of self-knowledge, Moran claims that "the *effort* involved in self-reflection, the struggle to get something right, and the characteristic risks of being wrong" (contra Bilgrami) all point to the objectivity of the phenomena of mental life (ibid., 40). The fact that we cannot simply bootstrap ourselves into more healthy or satisfying interpretations of ourselves (excepting instances of self-deception or delusion) indicates that "one's reflection is answerable to the facts about oneself, that one is open to the normal epistemic risks of error, blindness, and confusion" (ibid.). Thus, the substantial epistemic achievement taken to be involved in self-knowledge claims bolsters a metaphysical realism about mental states.

Moran's shift in focus places him as a transitional figure in my discussion. On the one hand, like Shoemaker, Burge, and Bilgrami he argues for an epistemic account of self-knowledge on the grounds of the connection it bears to rational agency. However, as I shall argue in Chapter 3, there is a tension between his argument for the substantiality of self-knowledge and what he sees as the unique commitment to first-order reasoning that defines genuine self-knowledge. While no doubt against his intentions, the emphasis on the importance of first-order reasoning in this regard may be read as potentially making room for an understanding of rationality and self-knowledge that need not appeal to any higher-order epistemic virtue on the part of the subject. For this reason I place him in Chapter 3, where I also consider Donald Davidson's non-epistemic account of self-

knowledge, or what he sometimes calls first-person authority. Here is a brief summary of Davidson's view.

1.7 Davidson: Semantic Authority Accounts for First-Person Authority

The thrust of Davidson's argument, found in essays such as 'First-Person Authority' (2001c) and 'Knowing One's Own Mind' (2001b), is that our knowledge of the beliefs and other propositional attitudes that we express through our sincere self-ascriptions is guaranteed by the fact that we cannot, generally speaking, fail to know the meaning of our words. He argues for this by combining two theses and an observation. The two theses are:

(1) the semantic externalist claim that the meaning of a person's words "depends in the most basic cases on the kinds of objects and events that have caused the person to hold the words to be applicable; similarly for what the person's thoughts are about" (Davidson: 2001b, 37);

and

(2) the regularity thesis, namely the claim that a subject is "not in a position to wonder whether she is generally using her own word to apply to the right objects and events, since whatever she regularly does apply them to gives her words the meaning they have" (ibid., 37-38),

The observation is that (i) as long as a speaker knows that she holds true the sentence she utters (i.e., is sincere), and (ii) knows what her words mean (as determined by the way

she consistently uses them), then she will know what she believes. With this, we can see how Davidson thinks the asymmetries between first- and third-person are explained. A speaker need not appeal to evidence, like others must, to know what she believes because the way in which she regularly uses her words constitutes their meanings (and thus the content of her belief as it is expressed through the use of those words). This is a guarantee she enjoys that her interpreter does not, for there is no guarantee that the use to which both put the words the speaker utters will be the same.

The special role that Moran assigns to first-order reasons points directly toward an expressivist account of self-knowledge, which Moran notes only to dismiss. As I will argue in Chapter 3, Davidson's strategy cannot succeed if not supplemented with an expressivist account of first-person authority that he never entertains. Both Moran and Davidson thus serve as transitional figures to the positive account of self-knowledge that I will recommend.

1.8 The Expressivist Account: Self-Ascriptions Express First-Order Mental States

The discussion and critique of Moran and Davidson in Chapter Three will show the need to consider a form of expressivism, which their views point towards, and which addresses the problems their views face. An appropriately nuanced expressivism asserts that the non-evidential basis and reliable truth of avowals is explained by the fact that such utterances ascribe the very beliefs they express. In other words, the essential claim is that, perhaps contrary to appearances, utterances of '*p*' and 'I believe that *p*' sometimes

express the same mental state of belief that *p*. However, it remains that, as indicated by their differing truth conditions, they mean different things. In other words, in the case of a special class of self-ascriptions, meaning and expressive content diverge. So the basic argument is this.[4] If my statement of 'I believe that *p*' serves to ascribe to me the belief that *p*, it follows that my *utterance* will be true if and only if I do in fact have that belief. But according to the expressivist thesis I also express the belief that *p*. Consequently, if I am sincere in my utterance of the self-ascription (i.e., I have the belief I express), then it follows that my utterance must be true. And this accounts for why, when I utter sincere self-ascriptions of my mental states, I will always get them right. In uttering 'I believe that Wagner died happy' I ascribe to myself (again, as indicated by its meaning) the very belief that my utterance expresses; assuming I am sincere, I will then have the belief I ascribe to myself. Thus, my sincere self-ascriptions will be true. The additional fact that they are expressions of mental states, and not assertions about them (that is, knowledge claims derived from some sort of cognitive act, for example some form of self-observation), explains why we can make them immediately and effortlessly, that is, without appeal to any evidence.

1.9 After Expressivism: Rationality Without Self-Knowledge?

If the expressivist explanation of self-ascriptive authority is sound, it would seem to pose a serious challenge to the idea of an essential connection between self-knowledge and

[4] The following synopsis derives from my reading of Jacobsen: 1996. See Bar-On (2004), Finkelstein (2000), and Hamilton (2000) for variations of this argument.

rational agency. After explaining the expressivist view I consider some recent criticisms of it made by Crispin Wright, Jane Heal, and Dorit Bar-On. Bar-On herself is sympathetic to the expressivist claim about the first-order expressive character of authoritative self-ascriptions. However, according to what she calls her "neo-expressivist" account, this does not rule out an epistemic understanding of those self-ascriptions. Instead, she argues for a "dual expressivist" understanding of self-ascriptions that takes them to express both the first-order state ascribed and the second-order belief that one has such a state. This account is motivated in part by what she takes to be valuable insights about the connection between self-knowledge and rational agency raised by, among others, Shoemaker, Burge, Bilgrami and Moran. Thus, after considering her view, I make use of David Owens' critique of epistemic agency to offer some reasons why self-knowledge cannot play the role in rational agency that they have seen for it. Finally, in concluding the discussion I briefly outline an alternative view of rationality – namely that of Donald Davidson – that offers an explanation of how it is that we may have rationality without self-knowledge.

1.10 Conclusion

I began this chapter by offering an introductory overview of some of the key issues that have recently been seen as defining the problem of self-knowledge. I then highlighted one prominent trend in recent discussions of self-knowledge that draws an explanatory link between our capacity for authoritative self-ascription of our mental states and our

status as rational agents. I also outlined a non-epistemic alternative to the explanation of self-ascriptive authority that would seem to undermine this general model. In the next chapter, I will provide a more detailed explanation and examination of the arguments of those philosophers who have argued for a particular kind of understanding of the link between self-knowledge and agency – what I have called the supervisory model of rationality (that is, Shoemaker, Burge, and Bilgrami). This in turn will set us up for the discussion of the non-epistemic option to come in Chapters 3 and 4.

Chapter 2: The Constitutive Relation Between Self-knowledge and Rationality

Introduction

In Chapter 1 I offered an overview of some key issues that have come to define the problem of self-knowledge. In the course of this discussion I pointed to a relatively recent trend amongst some philosophers to connect the explanation of the special features of self-knowledge to our status as rational agents. A sub-group, whose members include Shoemaker, Burge, and Bilgrami, explicates this relation in terms of the supervisory role second-order belief plays in the maintenance of rationality. In this chapter I take a closer look at each of these philosopher's views. I argue that, questions of the general plausibility of the supervisory model aside (this will be addressed in Chapter 4), each comes up short as a satisfactory understanding of authoritative self-ascription.

2.1 Sydney Shoemaker: The Rational Necessity of Self-Knowledge by Self-Acquaintance

2.1.a The Anti-Cartesian Thesis and Self-Blindness

In 'On Knowing One's Own Mind' (1996a) Shoemaker argues for what he calls a moderate Cartesian thesis: that the direct or privileged access to our mental states that we are said to have – what he calls self-acquaintance – is not a contingent fact about us, but rather is essential to the kind of mentality we enjoy. More specifically, he argues for a

constitutive link between first-and second-order beliefs that is directly tied to rationality – for a being with the kind of rational and conceptual resources we possess, having a first-order mental state that *p* necessarily entails that it have the belief that it has that state *p*.

He sets up the discussion by introducing the anti-Cartesian. The anti-Cartesian is one who thinks that our capacity for self-knowledge by self-acquaintance is like that of a hypothetical person who, after some training, is able to immediately report her blood pressure (Shoemaker: 1996a, 27). This is a "quasi-perceptual" capacity, in that the person's ability to do so at any given moment is logically independent of her having the blood pressure that she reports. For Shoemaker's anti-Cartesian, the situation is essentially the same with respect to our ability immediately and reliably to report our own mental states. While it may be that we can report our states in this manner, this capacity is not essential to having the states reported – as with the imagined blood pressure reporters, it is logically possible that we could have the latter without the former.

Shoemaker argues against this quasi-perceptual model of self-knowledge by arguing for the conceptual impossibility of something he calls self-blindness. Like a normal person (that is, one who has self-knowledge by self-acquaintance), a self-blind person is able to conceive of, and ascribe to herself the normal range of mental states; however, unlike someone who has self-knowledge by self-acquaintance, she has no way of determining the truth of those self-ascriptions (no way of knowing that she has the states she self-ascribes) except by third-person means (ibid., 30-31). Only if such a person were conceptually possible would it make sense to suppose that self-knowledge by self-acquaintance is a contingent fact about us rather than an essential component of

mentality. Hence, if it can be shown that self-blindness is conceptually impossible, the necessity of self-knowledge to mentality would be shown and the perceptual/empirical model of self-knowledge refuted.

2.1.b Moore's Paradox and the Argument Against Self-Blindness

According to Shoemaker the plausibility of the self-blindness thesis depends upon there being some sort of evidence for it in the form of behaviour that would distinguish the self-blind person from a normal person. This is based on the idea that if everything is as if a person has self-knowledge – that is, if she behaves just as a normal person would – then it is reasonable to conclude that she does have it (ibid., 36). To make her case, then, the anti-Cartesian would have to point to a situation where this sort of evidence would arise. Shoemaker begins the consideration of this possibility by discussing the role self-knowledge might reasonably be thought to play in critical rationality. That is, it might be argued that the rationalisation of the modification of first-order mental states requires second-order beliefs and desires: a desire to maintain coherence among one's beliefs and desires, a belief that certain first-order states are inconsistent with one another (which of course would require beliefs about what those first-order states are), and beliefs about what changes in those beliefs and desires would achieve coherence. On this supervisory model of rationality, the rationality (at a minimum, consistency and coherence) of a subject is secured through the oversight exercised by second-order mental states. The supervisory role played by our second-order states requires that we have knowledge of

our first-order states, and so self-knowledge is directly implicated in our rationality. If this is so, then of course self-blindness will be impossible. However, Shoemaker allows that here the anti-Cartesian might plausibly argue for the possibility of a person who, in the light of new experience that conflicted with her current beliefs and desires, was "hardwired" to make the necessary adjustments to those first-order states so as to maintain their coherence (ibid., 34). If so, then the person's lack of self-knowledge would show itself in her unwillingness to self-ascribe the relevant states (except on third-person evidence) as a normal person would. And this, it is argued, would prove the conceptual possibility of self-blindness and thus the contingent nature of self-knowledge by self-acquaintance.

In response to this Shoemaker offers what he calls the argument from Moore's paradox – a set of considerations designed to show that the putative self-blind person would be indistinguishable in his behaviour from a normal person, and that therefore self-blindness is conceptually impossible (ibid., 34 ff.). The idea here is that if a supposedly self-blind person could behave just like a normal person who is presumably not self-blind, that would reduce to absurdity the idea that the supposedly self-blind person was actually self-blind. He begins with the observation that on first inspection it might seem plausible that self-blindness would be revealed in the form of "Moore paradoxical" utterances – that is, those of the form '*P*, but I don't believe that *p*'. Given that the self-blind person (following Shoemaker we shall call him George) has otherwise normal cognitive and conceptual abilities, he should be capable of having and expressing beliefs about his environment. However, he should also be capable of forming third-person

beliefs about his own mental states. Shoemaker notes that it may be that the total evidence available to George would support the claim that, for example, it is raining. However, it may also be that the total "third-person" evidence – that concerning his behaviour – would support the belief that he does not believe that it is raining. In such a situation, he would be led on reasonable grounds to assert the Moore-paradoxical sentence "It is raining, but I do not believe it is raining," the utterance of which would distinguish him from a normal person and reveal his self-blindness.

Or would he? Shoemaker notes that, being as cognitively and conceptually able as a normal person, George would recognise the paradoxical, self-defeating nature of such utterances and would therefore avoid making them (ibid., 36). Presumably this would be grounded on his understanding of the nature of assertion and its connection to belief. Furthermore, he argues, it is reasonable to presume that the same ordinary cognitive and conceptual capacity that would prevent him from making such utterances would also have certain other effects that would make him indistinguishable from a normal person. For example, he would treat a question regarding his belief on a given matter ("Do you believe that *p*?") as questions regarding the truth of the matter ("Is that *p* true") and so answer appropriately (in other words, as a normal person would). In addition, he would understand that the meaning of "believe" would entail its use as a kind of assertion sign, but he would also have a grasp of the various considerations (Gricean and others, for example, when one meant to express uncertainty with respect to the assertion being made) that would determine when it was or was not appropriate to include it in one's utterances. Without going into further detail, the bottom line would seem to be that

George would behave just as a normal person would, thus undermining the idea that he could be self-blind.

Not so fast, the anti-Cartesian might protest. According to the above argument, George would be capable of avoiding the Moore-paradoxical utterances on the strength of his rational and conceptual capacities. Through them he could recognise the appropriateness of, and act upon, the rule: "If you have the intentions that make appropriate an assertive utterance of '*p*' do not conjoin this with an assertive utterance of 'I don't believe that *p*'". But is that so? George may recognise the rational force of the rule; the problem is, being self-blind, he would not know that he had the intentions referred to in the antecedent of the conditional, and would therefore have no reason to avoid the paradoxical utterance (thus revealing his self-blindness).

As Shoemaker points out, there is a problem with this reply. It is argued that George's lack of second-order belief regarding the intention that would motivate an assertive utterance of '*p*' would prevent him from following the rule that would allow him to avoid Moore-paradoxical utterances. However, the possibility of making a Moore-paradoxical utterance presupposes that he can assert '*p*'. But this would seem to be ruled out by the previous objection. If, to follow a rule of the sort "If you are in circumstances C, do (or don't do) X," one must recognise that one is in circumstances C (in this case, has the intentions that entail uttering '*p*' assertively), then self-blindness would rule out the assertive use of language (ibid., 36-37).

To this Shoemaker supposes that the anti-Cartesian would respond that this general claim about rule following is mistaken. The workings of a thermometer may be captured

by a rule ("when the temperature is X degrees Celsius, register X degrees Celsius") in spite of the fact that it does not recognise anything. Why, it may be asked, couldn't the following of certain linguistic rules be like this? In the case of assertion, a learning process would establish causal connections between certain intentional states and linguistic behaviour (assertive utterance). If so, then there would be no need for the speaker to recognise, in the form of true second-order beliefs, that she had those states in order to make those utterances. If this were true, then the anti-Cartesian's claim regarding the possibility of Moore-paradoxical utterances that would reveal George's self-blindness might be preserved. This would depend on the same observations not applying to the kind of rules required to avoid Moore-paradoxical utterances (ibid., 37).

In response to this Shoemaker argues that, given that George has normal cognitive and conceptual abilities, we have no reason to think that they would not:

> If despite his self-blindness George could acquire the assertive use of language, then in doing so he would also learn to use "believe" in such a way as to avoid pragmatic paradox, and what goes with this, to give appropriate answers to questions of the form "Do you believe that *P*?", to preface certain kinds of assertions with the word "I believe," and in general to be indistinguishable from someone having the faculty of self-acquaintance." (ibid., 37-38).

The upshot of these considerations is that the anti-Cartesian must abandon the Moore's paradox gambit. However, it does not follow that self-blindness is ruled out. For as it stands, these considerations only show that a self-blind person could not make

assertions (unless he acquired the requisite second-order belief through third-person means).

To answer this objection Shoemaker argues as follows. Given that George has normal intelligence and conceptual capacity, he should be able to come to understand language. But if this is so, he should be capable of employing this knowledge in learning to make assertions. Given his rationality, he could see that various goals of his would be met by making certain utterances, which would lead him to say those things, just as in other cases he is moved to do what he believes will promote the realisation of his goals. With respect to assertion, he could come to see that uttering certain sentences will have the effect of promoting in the hearer a belief in the proposition uttered. So, with this goal in mind, he would be led to make such assertive utterances (ibid.).

According to Shoemaker, the availability to George of this sort of reasoning would account for a great many of the cases where assertion would arise (ibid., 40). However, it may be objected that one sort of case would be left out, namely where the speaker would be motivated by an intention to tell another about his beliefs about a given matter. To illustrate Shoemaker offers the following scenario: A person – call her Anna – is seeking a lucrative partnership with a fellow *p*-believer. George is introduced to Anna and realises that it would be to his advantage if he were to form a partnership with her, assuming of course that he is also a *p*-believer. If, like a normal person, George had self-knowledge by self-acquaintance and were a *p*-believer, he would simply inform her of his status by saying '*p*' or 'I believe that *p*'. Unfortunately, lacking such self-knowledge, he would have to remain mute (this presupposes that there would be no third-person

evidence upon which he could draw to come to a conclusion regarding his belief). And this, it could be argued, would serve as evidence for his self-blindness.

However, as in the other cases, Shoemaker argues that George would have an argument open to him that would compensate for this apparent deficiency (I will call this the "argument for reporting"). He could reason:

(1) '*p*' is true. (As Shoemaker notes, this states his belief, but doesn't say that he has it, so presumably he remains in the dark regarding whether or not he believes it.)

(2) Generally speaking it would be to anyone's advantage to act on the assumption that *p* is true, for acting on true assumptions usually furthers one's ends.

(3) This entails that I ought to act on the assumption that *p* is true.

(4) This means that I ought to act as if I believed that *p* is true, which would include saying '*p*' or 'I believe that *p*'.

(5) This would have good consequences for me, as (i) it would lead Anna to choose me as her partner, and (ii) a team of persons acting on the assumption of *p* – something true – would likely be successful in their endeavours.

(6) This would provide George with a motive for saying '*p*' or 'I believe that *p*'.

Thus, once again, it is argued that George would act just like a normal person. And, given any lack of distinguishing behaviour, there would be no reason to suppose that he was self-blind (ibid., 40-41).

2.1.c Self-Knowledge Via Self-Intimation

From these and other similar considerations (Shoemaker provides a brief argument to show that what has been argued for belief applies to other mental states that factor in to the rationalisation of behaviour) we find that George would be indistinguishable in his behaviour from a normal person (someone who has self-knowledge by self-acquaintance). Shoemaker thinks this supports two related conclusions. One is that self-blindness is conceptually impossible, and thus self-knowledge by self-acquaintance is essential to mentality. The other is that our capacity for such self-knowledge is best explained by the self-intimating character of first-order mental states.

Before explaining how it supports the latter two terms need to be introduced. The first is "mental assent": when a thought occurs to one (for example, that it is raining), one may assent to it (that is, endorse it as true), or not. According to Shoemaker, assent is an "episodic instantiation of belief" (Shoemaker: 1996d, 78). The second term is "available belief": an available belief is one that "is apt to serve as a guide to action, and what goes with this, is apt to be among the premises of the subject's reasoning"(ibid., 80).

The self-intimation thesis has two parts:

(1) If a belief is available, and it is presented as a candidate for assent, then the subject will assent to it.

(2) If a belief that *p* is available to the subject, then she has the tacit belief that she has that belief, and that second-order belief is available as well.

By (1) it follows that, if one assents to '*p*', then (i) one believes it and (ii) that belief was available to the subject. If so, then by (2) the subject will at least tacitly believe that she

believes that *p*, and that second-order belief will also be available to her, to which, if it is presented as a candidate for assent, she will assent as well.

Shoemaker maintains that (1) needs no supporting argumentation, as it is partly definitive of an available belief that one would assent to it if it were to come up for consideration. However (2) does need support, and this is to be found in the 'argument for reporting' outlined above (see 2.1d). That argument was said to show that a rational subject who believed that *p* would have, and be capable of employing, a line of reasoning that in the relevant circumstances would dispose her to behave in ways that (i) would indicate possession of an available belief that she believes that *p*, and (ii) would manifest assent to that proposition (for example, in the appropriate situation she would be disposed to offer linguistic assent to it in the form of the assertion of "I believe that *p*," or in her disposition to answer affirmatively to "Do you believe that p?"). Given that this line of reasoning that would so dispose her to behave would be available only if the belief that *p* were available to her, it follows that a subject who has the available belief that *p* has at least the tacit available second-order belief that she believes that *p*. This, Shoemaker contends, constitutes strong support for the self-intimation thesis and the idea that self-knowledge by self-acquaintance is essential to mentality (where the knowledgeable status of the second-order beliefs derives from the fact that they are reliably produced, and we are entitled to see them as such).[5]

[5] Shoemaker writes: "The higher-order beliefs from which this issues count as knowledge for a familiar reason; they are reliably produced, and we are entitled to regard them as reliably produced." (Shoemaker: 1996b, 92)

2.1.d The Argument for Reporting and the Assumption of Second-Order Belief

To summarise, Shoemaker argues that in every case in which we might suppose George's self-blindness to manifest itself it turns out that he would have a line of reasoning open to him that would dispose him to act just as a normal person with self-knowledge by self-acquaintance would. There are two possible conclusions to be drawn from this. One is that the supposedly self-blind person could not be self-blind after all – since everything in his behaviour would be as if he has self-knowledge by self-acquaintance, there would be no good reason to suppose that the putatively self-blind person would actually be self-blind. Rather, given that George's possession of first-order beliefs and desires plus normal rationality and conceptual resources suffices for the explanation of his behaviour, we ought to see the second-order beliefs constitutive of our self-knowledge as supervening on these properties – that, as Shoemaker puts it, having the former is nothing over and above having the latter (Shoemaker: 1996a, 34). For "there is no phenomenology of such states that is in danger of being ignored if we say this – there is nothing it is like to believe something, and there need not be anything it is like to know or believe that one believes something" (ibid.).

In 'Self-Knowledge and Inner Sense' (1996c) Shoemaker remarks that one might find fault with this account of self-knowledge. It has been argued that George, in virtue of his following the line of reasoning outlined above, would act just as a normal person would. But, it might be countered, we, who have introspective self-knowledge, do not *use* such lines of reasoning to self-ascribe our first-order mental states. Furthermore, if it were the

case that George could only behave in a way consistent with having second-order beliefs in virtue of employing such reasoning, it would entail that in fact he was self-blind after all, even if nothing would be revealed by his behaviour (Shoemaker: 1996c, 239). These worries, Shoemaker says, are based on a misunderstanding of the intention of the argument. To point out that the reasoning outlined above is available to the supposed self-blind person is not to suggest that we normally engage in it – obviously, he says, we don't. Rather, it is meant to point out that,

> ...in order to explain the behaviour we take as showing that people have certain higher-order beliefs, beliefs about their first-order beliefs, we do not need to attribute to them anything beyond what is needed in order to give them first-order beliefs plus normal intelligence, rationality, and conceptual capacity. What the availability of the reasoning shows is that the first-order states rationalise the behaviour. And in supposing that a creature is rational, what one is supposing is that it is such that its being in certain states tends to result in effects, behaviour, or other internal states, that are rationalised by those states. Sometimes this requires going through a process of reasoning in which one gets from one proposition to another by a series of steps, and where special reasoning skills are involved. But usually it does not require this. (Ibid.)

In other words, the point here is not to argue that we normally engage in the kind of reasoning outlined above. Instead, it is to point out that that since the first-order states plus rationality and conceptual capacity are all that is needed to explain the second-order beliefs, it follows that those second-order beliefs are "generated" by them. But we

needn't engage in such reasoning for the effects – the second-order beliefs – to be realised.

That said, one might argue that this question, and Shoemaker's response to it, don't quite get to the heart of the matter. His reply seems to be that while we need not, and don't normally, employ the kind of reasoning mentioned above (what I have called the argument for reporting), it may still be cited in the rationalisation of what we take to be second-order behaviour, and that consequently this lack of employment does not refute the self-intimation thesis or the constitutive relation between first-order states, rationality, and self-knowledge. But – and this brings us to the second possible conclusion alluded to above – why should one not suppose that since, as Shoemaker says, "normal rationality and intelligence plus first-order beliefs and desires gives you everything in the way of explanation of behaviour that second-order beliefs can give you" (Shoemaker: 1996a, 48), in fact such second-order beliefs are superfluous and the anti-Cartesian is correct after all? Shoemaker recognises this possibility, but notes that this would be to conclude that self-acquaintance confers no advantage on one who has it, and would even raise doubts as to whether we possess it at all. As he says, it would even be conceivable that self-blindness was the normal condition of humanity. However, rather than conclude this, he argues that we ought to take the implausibility of such consequences as entailing a *reductio ad absurdum* of the possibility of self-blindness.

This might strike one as a somewhat curious conclusion to draw. On the one hand, by his own account, Shoemaker takes himself to have shown that, at the very least, second-order beliefs are not necessary for the rationalisation of normal human behaviour.

Furthermore, as he sees it, there would be no phenomenological difference in the mental life between one who had, vs. one who lacked, second-order belief by self-acquaintance (recalling that, as he points out, one need not go through an explicit process of reasoning for one's states to be rationalised, which would apply whether or not the behaviour rationalised involved first- or second-order beliefs). On the other hand, he argues that the lack of difference in the behaviour of the self-blind and normal persons reduces to absurdity the idea that self-blindness is a conceptual possibility. So, on what does the *reductio* claim rest?

A couple of apparently innocuous assumptions seem to underlie Shoemaker's position here. One is that certain behaviour as it exists in us – for example, deliberation (and the notion of agency that goes with it), and the communication of one's mental states, especially in the form of self-ascriptions – obviously involves introspectively acquired second-order belief. Indeed, the self-intimation thesis seems to rest on it, as it is invoked to explain how behaviour taken to manifest second-order beliefs can arise from the combined forces of our first-order states, rationality, and conceptual abilities. The second general assumption, consistent with the self-intimation thesis, is that awareness of one's first-order states requires second-order beliefs about them. The implication seems to be that if self-blindness were a conceptual possibility in spite of the fact that the self-blind person would be indistinguishable in his behaviour from us, then certain obvious facts about our mental lives – that we are intentional agents, that normally we know our own minds and can authoritatively report on them – would be reasonably challenged.

And this, Shoemaker suggests, is absurd, and makes the self-intimation thesis the more reasonable conclusion.

However, there may be another way to look at the matter. One might argue that if the phenomena that Shoemaker associates with self-knowledge by self-acquaintance can be explained without appeal to second-order belief (for example, by our ability to be immediately aware of and authoritatively report on our own mental states, or our capacity for critical deliberation about what to believe or desire), then the anti-Cartesian claim may come to seem less absurd. In Chapters Three and Four I shall explore the plausibility of this idea.

2.2 Tyler Burge: Self-Knowledge and Critical Rationality

2.2.a Anti-Individualism and Self-knowledge: the Sceptical Worry

Tyler Burge's initial discussion of self-knowledge takes place against the backdrop of his semantic externalist (or what he calls anti-individualistic) theory of thought content. There are two forms of anti-individualism for which he argues: (i) perceptual externalism, and (ii) social externalism. As he states it in 'Individualism and Self-Knowledge' (1998b), "individuating many of a person or animal's mental kinds ... is necessarily dependent on relations that the person bears to the physical, or in some cases social, environment" (Burge: 1998b, 112). Some, such as Donald Davidson, have argued that anti-individualism leads to skepticism regarding the possibility of privileged knowledge of our own thoughts (Davidson: 2001b). The concern is that if the

individuation conditions of the content clause in a present-tense indicative self-ascription of a thought that *p* (what Burge terms basic self-knowledge [Burge: 1998b, 112]) depend upon relations to the social or physical environment of which one might be ignorant, then it becomes questionable that one knows what one is thinking when one makes such a self-ascription. According to Burge this worry is misplaced, since the positive epistemic status of such self-ascribed thoughts does not depend upon the possession of knowledge of causal/perceptual relations the thinker has to her environment (or, presumably, to the communal linguistic standards to which one holds oneself responsible), but rather derives from their reflexive, self-verifying character.

2.2.b Burge's Social Anti-Individualism

According to Burge's social anti-individualism, the content of one's words and propositional attitudes is determined in part by the communal standards to which one defers (that is, to which one holds oneself responsible) in one's linguistic usage. To illustrate he offers the following thought experiment (Burge: 1998a, 26 ff.). First, we are asked to imagine a person who is "generally competent in English, rational and intelligent". Through "casual conversation or reading" he has acquired a wide variety of true beliefs about arthritis that are reflected in his use of the term. However, never hearing or reading anything that would suggest either that arthritis could or could not occur in the thigh, he comes to just such a belief about himself. He then visits a doctor to whom he conveys his belief that he has gotten arthritis in his thigh. The doctor responds

by telling him that that cannot be so, since (as any dictionary would have indicated) arthritis is a disease of the joints only, so he must have some other ailment afflicting his leg. After expressing surprise the patient readily gives up his belief and wonders what, then, might be the problem with his thigh.

Next Burge asks us to consider the following counterfactual situation. Everything in terms of the patient's history, physical make-up, behaviour and dispositions to behave, and phenomenal/internal qualitative states remains the same. However, in this case, the standard use of the term 'arthritis', as it is both conventionally and defined to apply, includes what is said to be the former misuse (his lack of error, one might say, is due to semantic luck). Both experts and informed laypersons apply 'arthritis', not only to arthritis as previously conceived, but also to various other rheumatoid ailments, including those that might occur in the thigh.

The thought experiment is meant to illustrate that mental content is not solely in the head, but is partly constituted out of relations the subject bears to his or her social linguistic community. While everything with respect to the patient considered in isolation from the standard linguistic usage of his community remains the same in the two situations, the patient expresses different beliefs by "I have arthritis in my thigh." In the actual world case, he expresses the false belief that he has arthritis in his thigh (where the extension of 'arthritis' is restricted to diseases that cause inflammation in the joints). However, due to a partial misunderstanding of the concept, he misapplies it, and so is mistaken. In the latter (counterfactual world) case, he expresses a true belief with a different content, one that includes ailments of the thigh (say, *tharthritis*). The difference

in mental content is attributable to a combination of two factors. One is the intention of the subject to speak as others in his speech community do; the other is the content of those linguistic norms. Generally speaking, subjects hold themselves (and each other) responsible to the linguistic or conceptual norms of their community.[6] Where this is reasonably seen as the case (for example, where a subject demonstrates general competence in the usage of the language in question, is of reasonable intelligence, does not insistently flout any of the conventions in question), it is correct to ascribe the meaning to her words she intends, in spite of her mistaken usage. To return to Burge's example, in each application of 'arthritis' the patient intends to speak in accord with the standards or norms of those in the linguistic community to which he defers (as indicated by his ready willingness to give up his belief that he has arthritis in his thigh), and thus should be so interpreted. Having otherwise demonstrated a general competence in his use of the term, this mistaken belief about what is defined and commonly understood to be arthritis in the actual case should not be seen as affecting the content of his mental state.

2.2.c Perceptual Anti-Individualism

In the case of Burge's perceptual anti-individualism the situation is similar, the main difference being the nature of the external factor that partly determines word meaning

[6] "All I rely on is the fact that individuals actually regard themselves as responsible to linguistic or conceptual norms that might be applied to them by others. This much seems implicit in the notion of interpersonal agreement and disagreement" (Burge: 1982, 291).

and mental content. To think 'water is a liquid' one must be a competent user of those words, the meaning and reference of which is determined in part by the physical environment of those who use them. Having 'water' thoughts partly depends on one having been in causal relations with water for long enough to set up a semantic connection between water and one's use of the word form 'water'. Word meaning and associated thought content are individuated in part by the nature of the entity with which one causally interacts.[7]

2.2.d Davidson's Critique: Burge's Anti-Individualism Undermines Authoritative Self-Knowledge

One might argue that these anti-individualist views of meaning and mental content are at odds with the idea that we enjoy a special kind of authority with respect to knowledge of our thoughts and other mental states. Recall Burge's patient, who, in spite of having all the same physical states, history, behavioural dispositions and inner qualitative or phenomenological experience, would think different thoughts depending upon in which linguistic community he resided. In the actual world, his thoughts as expressed through his use of 'arthritis' would be about arthritis; if he were transferred to the counterfactual world, they would be about tharthritis. The situation is much the same with regard to perceptual externalism. To borrow Burge's example, say that, unbeknownst to me, a "mischievous genius" switches me from Earth to Twearth at long enough intervals for me

[7] "Let us assume that our thoughts about the environment are what they are because of the nature of the entities to which those thoughts are causally linked" (Burge: 1998b, 114).

to interact with water (H_2O) and twater (A_2Z) so that both concepts become part of my conceptual repertoire (Burge:1998b, 115 ff.; 1998c, 243). Since I am identical in all the ways outlined above, upon being switched I would have no reason to suppose that my thought content had changed. It would seem to follow, as Davidson points out, that in such situations I could not know what I believed (Davidson: 2001b, p. 26).[8]

That one might be mistaken about or only partially understand one's mental states is not restricted to such counterfactual or switching scenarios. As Burge notes, on his view the phenomena of partial understanding, or even misunderstanding, of the meanings persons use to express their mental states are not unusual. If so, it would seem that speakers are not as authoritative about what they mean, and therefore what they think, believe, desire, or intend as expressed by their words, as is commonly thought. For if I don't fully grasp what a given word by which I express my mental state means, then it would appear that in some respect I do not know in what the content of that mental state consists. Davidson takes this to show that, on Burge's view, "first-person authority is very seriously compromised" (ibid., 22). Burge disagrees – he denies that the partial understanding that must be admitted if his version of externalism is to stand entails any loss of what we normally take to be self-knowledge.

2.2.e Burge's Reply to Davidson

[8] Davidson: 2001b, 26. Davidson makes this point within his discussion of Burge's social externalism, but it applies to the perceptual case as well.

Burge speculates that Davidson may be led to this conclusion by his failure to appreciate the distinction between "'knowing what one's thoughts are' in the sense of basic self-knowledge [and] knowing what one's thoughts are in the sense of being able to explicate them correctly – delineate their constitutive relations to other thoughts" (Burge: 1998b, 125).[9] The former (being able to self-ascribe one's thoughts), he argues, does not require the latter. Thinking that *p* (say, that one has arthritis or that water is a liquid) requires that certain conditions be met (for example, that one be generally competent in the use of the terms in question and bears certain relations to a particular social and physical environment). Knowing what we think, in the sense of basic self-knowledge, requires only that we be able to think such a thought self-ascriptively. Due to their reflexive nature, such self-ascriptive judgements are self-verifying – in the process of judging that one thinks that *p*, one also thinks that *p*, or the very thought that one judges oneself to be thinking. Their self-verifying character accounts for the truth and justification of these judgements.

Again, one need not know that the individuating conditions needed to think the thought that *p* are in place to think that thought; one also need not know that those conditions are in place to make the second-order judgement that one is thinking such a thought. One might be tempted, Burge suggests, to think that the justification of one's second-order belief requires such knowledge; we adopt a third-person perspective and imagine that the world may be other than we suppose, and that our thoughts might differ

[9] Actually, this does not seem to be the focus of Davidson's objection; the problem is not that one would be unable to explicate one's thoughts in terms of, for example, their inferential connections to others (but this might not be Burge's idea here), but rather that one might remain ignorant of key aspects of the content of the mental state one self-ascribes.

accordingly. We are then seduced into the belief that unless such a possibility can be ruled out the self-ascription is insufficiently justified (ibid., 124). In response to this Burge draws a comparison with epistemic norms associated with perceptual-based knowledge claims. We don't need knowledge of these epistemic norms to make perceptual judgements that count as knowledge. Generally, entitlement does not depend on the possession of a set of independently obtained justified true beliefs regarding the presence of the necessary preconditions that make the perceptual judgement possible. Generally speaking, perceptual knowledge that there is a barn in front of one does not require prior knowledge that all the enabling conditions for such knowledge obtain (such as appropriate lighting conditions, or a lack of cunningly realistic barn facades in the vicinity). But if it is not required in this type of knowledge, why, in effect he asks, ought it be required in the case of one's knowledge of one's mental states (ibid., 117)?[10]

2.2.f Burge and Davidson's Competing Views of Meaning and Mental Content

Even if we agree with Burge here (as does Davidson), this does not address the problem, as Davidson sees it, of partial understanding of the criteria needed to think a thought that *p*. The difference between the two rests on their competing views regarding the nature of

[10] One's first response might be to agree – in normal cases of empirical knowledge claims we don't and ought not demand such background knowledge – but then add that the slow-switching scenario set up to illustrate the original semantic claim is not a normal situation. Given the context, where one is aware that one possesses phenomenologically indistinguishable thoughts ('water' and 'twater'), it would seem reasonable for one to require evidence that one is on Earth or Twearth in order to know (provide adequate justification) that one is thinking Earth- or Twearth-based thoughts. However, according to Burge this would be to mistakenly assimilate basic self-knowledge to 'explicatory' self-knowledge (that is, the knowledge needed to fully explicate a given concept) (Burge: 1998b, 125).

meaning and mental content. There are two key differences. One concerns the role external norms are thought to play (or not) in the constitution of meaning and mental content. The other concerns the extent to which the content of one mental state is dependent on the content of others. With respect to the former, Burge argues that the meaning of the words a subject uses to give content to her propositional attitudes is partly determined by external norms, the nature of which provide the criteria for the correct application of those words. For Davidson, the order is reversed – he argues that how one regularly uses one's terms, including whatever (in the most basic cases) one regularly applies a term to, gives those words the meanings they have (Davidson: 2001b, 37). As he sees it, there is no external norm the subject need grasp (no object of thought); discernible regularity in the application of a term is the only norm that is needed to determine meaning.[11]

So for Burge, the correct use of the term 'arthritis' is determined by the meaning of that term as determined by the nature of the relevant linguistic community's standard use. One may fail to properly grasp the content of the external norm as, for example, when she applies it to an ailment in her thigh, but this does not necessarily affect the content of her mental state as determined by the nature of the external component. In spite of her idiosyncratic use, she should be interpreted as having a belief with the same content as any other member of her community would have, given his or her correct use of the term. On Davidson's view, since she regularly applies the term 'arthritis' to afflictions of the thigh, that use ought to be included in the meaning assigned to her word and associated

[11] This is somewhat redundant, as it is the presence of a discernible regularity that makes the utterance of a term a use, i.e., an application that can be considered correct or incorrect.

mental content, irrespective of how others in her community speak (or of her intention to speak as those others do). Thus, for Burge, her utterance of 'I have arthritis in my thigh' will be false, while for Davidson it will be true.

This difference in their positions regarding the nature of meaning helps explain the dispute regarding the nature and extent of a subject's authority regarding the content of her mental states. Burge's distinction between basic self-knowledge and what I have called explicatory self-knowledge is a consequence of his version of semantic externalism. If what one means is determined in part by relations to external factors of the kind Burge envisions, conceiving of self-knowledge along explicatory lines would undermine the idea that we enjoy a unique kind of status with regard to knowledge of our mental contents. For on Burge's view, we do not enjoy any special "authority about whether one of one's thoughts is to be explicated or individuated in such and such a way" (Burge: 1998b, 125). According to Davidson, it is just the opposite. Since it is regular use that determines meaning, such authority is unavoidable. As he says, on his view "nothing could count as someone regularly misapplying her own words" (Davidson: 2001b, 38).

2.2.g Burge's Account of Our Entitlement to Non-Basic Self-Knowledge

We will return to the above matter in Chapter 3. For now we can observe that for Burge, in the case of basic self-knowledge, a subject's authority with regard to her self-ascriptions derives from their contextually self-verifying nature. They are groundless, in the sense that they are "based on nothing" (or "nothing empirical," as Boghossian

expresses it [Boghossian, 156]) save an ability to think the given thoughts in a second-order way. Burge argues that this ability is sufficient for the truth and justification of this class of second-order judgements. However, as he also notes, this class of self-ascriptions demarcates a fairly narrow subset of the full range of self-ascriptions that are thought to count as privileged knowledge (Burge: 1998c, 241). Most of our self-ascriptions, for example of our beliefs, desires and intentions, are not of the contextually self-verifying kind, and require a different explanation for the subject's authority in making them. Furthermore, as Boghossian points out, an account of that authority should make room for the possibility of error (Boghossian, 151). Normally we do not take self-ascriptions to be infallible; they are defeasible, subject to rejection if evidence indicates otherwise. But, according to Burge, basic self-knowledge is, save extreme cases of "cognitive pathology," infallible (Burge: 1998c, 241). However, in most cases I may self-ascribe a belief that *p* with understanding, but that alone does not guarantee that I have the belief I supposedly judge myself to have. For these other sorts of self-ascriptions another account will be needed. According to Burge, this is to be found in an understanding of the role they play in critical rationality.[12]

2.2.h Why Our Second-Order Beliefs Must Count as Knowledge

Burge offers a transcendental argument to explain why our non-*contextually self-*

[12] In fact Burge mentions that his description of the contextually self-verifying character of the self-ascriptions classified as basic self-knowledge does not exhaust their epistemic status; all self-ascriptions, he argues, share in the entitlement that the link to critical reason supplies (Burge: 1998c, 242, 245).

verifying, non-self-interpretive second-order judgements must be knowledgeable.[13] The anchoring phenomenon is critical reason: "All of us," Burge says, "even sceptics among us, recognise a practice of critical reasoning" (ibid., 246). The basic claims are that self-ascriptive judgements play an essential role in critical reason, and that critical reasoning requires that such judgements be knowledgeably made. Regarding the former, Burge writes:

> As a critical reasoner, one not only reasons. One recognises one's reasons as reasons. One evaluates, checks, weighs, criticizes, supplements one's reasons and reasoning. Clearly this requires a second-order ability to think about thought contents or propositions and rational relations among them. (Ibid.)[14]

This is not to say that all reasoning by critical reasoners is of this kind. Still, for a critical reasoner to be a critical reasoner, this standpoint must always be available. But why must our second-order judgements – including our second-order beliefs about the content of our first-order mental states – count as knowledge?

The status of our second-order judgements as knowledge requires that they be both justified and true. Looking first at the issue of epistemic warrant, Burge begins by arguing that there is more to that idea than the ordinary notion of justification. An individual's epistemic warrant for a given judgement may derive from "an *entitlement* that consists in a status of operating in accord with norms of reason, even when those

---

[13] By self-interpretive self-ascriptions I mean those derived from one's interpretation of one's thoughts, emotions, or behaviour.

[14] Burge adds that this reasoning is not merely about propositions and their entailment relations. Since critical reason involves the assessment of the truth and reasonability of reasoning, its focus is not merely on the content of the propositional attitudes, but the attitudes themselves (Burge: 1998c , 247).

norms cannot be articulated by the individual who has that status" (ibid., 241). This, he asserts, is the type of epistemic warrant that applies to our privileged self-knowledge. To see how and why our second-order beliefs are warranted we need only understand how the ability to make correct judgements about one's mental states makes an essential contribution to the rationality of the subject. Our second-order judgements about our first-order mental states derive their warrant from the supervisory role they play in the process of maintaining rational coherence in one's mental life. Critical reason involves making judgements about the content of one's first-order mental states as well as the rational assessment of those states, the purpose of which is the promotion of general reasonability in the subject. If one were unreasonable in, and hence lacked entitlement to, one's second-order judgements, then the rational connection between those judgements and the first-order mental states they are meant to assess and guide would dissolve. If such judgements were not themselves reasonable – in accord with the norms of reason – they could not serve their function of promoting the overall rationality of the subject. Thus, the sort of second-order judgements involved in the process of critical reason are by their nature epistemically warranted.

The argument for the second half of the epistemic equation follows along similar lines. Our second-order judgements must be generally veridical because systematic error regarding the content of first-order mental states would undermine our entitlement to those judgements. If our judgements about the content of our mental states were not generally true, then any second-order or reflective reasoning involving them could not guide the subject in her evaluation and control of her first-order states that is essential to

promoting the general reasonability of the subject. But, as Burge has argued, this is just what critical reason enables. Thus, so long as a subject is interpretable as generally reasonable, and we understand the role of critical rationality in this reasonability, we have an explanation of why our second-order beliefs about our first-order states must generally be considered epistemically warranted and true – that is, knowledge. As Burge puts it, "understanding and making such judgments is constitutively associated with being reasonable and with getting them right" (ibid., 245). Understanding and making correct second-order judgements is necessary for being reasonable – we wouldn't be (full-blown) reasoners without these capacities. At the same time, their reasonableness – that they accord with norms of reason – provides their epistemic warrant. So by their nature such judgements are normally veridical and epistemically warranted.

2.2.i Burge's Non-Empirical Substantial Epistemology of Self-Knowledge

In linking our epistemic entitlement to non-contextually self-verifying self-ascriptions to critical reason Burge offers what could be described as a "weak" or non-empirical, substantial epistemic account of self-knowledge.[15] Compare his account of basic self-

[15] A weak or non-empirical substantial epistemic account argues for what Elizabeth Fricker calls "weak special access". While there is no substantial cognitive achievement in Boghossian's sense of the term (that is, knowledge gained through inference or observation [Boghossian, 165]), a form of special access remains in that the second-order belief "tracks" an ontologically distinct first-order state. As Fricker puts it, "the core idea [is that] of causally mediated reliable tracking by self-ascriptive beliefs of ontologically distinct first-level mental states which they are about" (Fricker, 161). This may be contrasted with what Boghossian calls an "insubstantial" epistemic account (Boghossian, 166), namely one that takes self-ascriptions to express second-order knowledge without any phenomenon of reliable tracking (as in the case of Burge's basic self-knowledge).

knowledge. In basic-self-knowledge, the reliable truth of self-ascriptions is explained by the ability to think thoughts in a second-order way – in thinking that one thinks that *p* one also thinks that *p*. As contextually self-verifying, it is thought and thought about at one and the same time. Such an account is insubstantial in the sense that it does not appeal to some form of reliable tracking to explain either how such second-order judgements arise, or why they are reliably true or warranted.

The situation is partly the same with respect to the critical reason account. On the one hand, Burge argues against there being any room for contingent causal relations in the explanation of our epistemic warrant for our non-contextually self-verifying self-ascriptive judgments (ibid., 245).[16] On the other hand, he does seem to acknowledge that first-order mental states and second-order beliefs about them are ontologically distinct states. Indeed, his argument for entitlement would seem to entail it. Burge argues that the ability to reason critically about one's first-order mental states is a necessary condition for the possibility of thinking first-person present-tense thoughts about those states. As he says, "I think that the following necessity holds: To think the relevant first-person present-tense thoughts about one's thoughts and attitudes, one must be capable of critical reasoning" (ibid., 246). While critical reasoning is not always focused on one's first-order intentional states as one's states, it does require the ability to make such judgements – the identification and evaluation of one's first-order mental states is one of the key functions of critical reason:

[16] In other words, contra Shoemaker, Burge contends that a causal-reliabilist account of warrant will not do as an explanation of how second-order belief counts as knowledge.

> Critical reasoning involves an ability not merely to assess truth, falsity, evidential support, entailment, and nonentailment among *propositions* or *thought contents*. It also involves an ability to assess the truth and reasonability of reasoning – hence *attitudes*. This is not to say that critical reasoning must focus on attitudes, as opposed to their subject matter.... But to be fully a critical reasoner one must be able to – and sometimes actually – identify, distinguish, evaluate propositions as asserted, denied, hypothesized, or merely considered. (ibid., p. 247)

Also, in arguing against what he calls the "simple observation model"[17] of epistemic warrant for our self-ascriptions he mentions that it is not the existence of a causal relation between a first-order mental state and second-order judgement about it that is at issue, but only the part such a relation might play in an explanation of one's entitlement to those second-order judgements. He writes:

> It is common to my view and the opposed observational view of self-knowledge that in many of the cases under dispute, there is a causal mechanism that relates attitudes to judgements about them. What is in dispute is the nature of the epistemic entitlement that one has to such judgements, not the existence of a psychological mechanism. (ibid., 254, n. 12)

Such a mechanism may be a "background enabling condition" for our entitlement, but it can never be part of the substantive explanation of why it is that we are so entitled.

---

[17] The simple observational model is described as follows. "The fundamental claim is that is that one's epistemic warrant for self-knowledge always rests partly on the existence of a pattern of veridical but brute, contingent, non-rational relations – which are plausibly always causal relations – between the subject matter (the attitudes under review) and the judgments about those attitudes"(Burge: 1998c, 253).

So to summarise, Burge offers a non-observational or non-empirical explanation of self-knowledge that is "weakly" substantial. While he denies that causal relations of the sort normally associated with observational models play any role in the explanation of epistemic warrant, he still maintains that self-ascriptions express second-order beliefs that are ontologically distinct from, and causally related to, the first-order mental states they are about.[18] I shall address the question of how plausible this approach may be shortly. But first, I wish to examine how and to what extent it explains the asymmetries that characterise self- and other-ascriptions of mental states.

2.2.j How does Burge's Account of Epistemic Entitlement Stand With Respect to the Asymmetries?

How does Burge's view of self-ascriptions, as entailed by his explanation of epistemic entitlement, stand with respect to the asymmetries associated with self- and other-ascriptive utterances? Take first the idea that persons appear to possess a level of authority with respect to certain of their self-ascriptions that they do not enjoy with regard to their attribution to others (or others to them). While Burge does not explicitly address this matter, the reasoning of his transcendental argument provides a possible explanation of this asymmetrical security in ascription. Given (i) the role of true self-ascriptive judgement and belief in critical reason, and (ii) critical reason in the maintenance of the overall rationality of the subject, and (iii) the necessity of such rationality to the interpretability of the subject, it follows that the possibility of being

[18] In this way his is consistent with Shoemaker's account.

interpretable is grounded in part upon one's normally getting right one's judgements about one's own mental states. A failure to do so would indicate a rational breakdown in the subject that would impede another's interpretation of her mental states. As Crispin Wright expresses it, "[w]holesale suspicion about my attitudinal avowals (self-ascriptions) – where it is not about sincerity or understanding – jars with conceiving of me as an intentional subject at all" (Wright: 2001d, 325).

What of the second component of our social epistemic practice, viz., the groundless character of self-ascriptions? In the case of basic self-knowledge this is explained in terms of the contextually self-verifying character of the self-ascriptions, where one thinks the first-order thought in the process of making the second-order judgement that self-ascribes it. There is no *way* in need of explanation regarding how one knows what one is thinking, other than the ability to think the thought in a second-order way. The self-evident, self-verifying character of the self-ascription obviates the need for such explanations.

What is the situation with non-contextually self-verifying self-ascriptions (i.e., those such as 'I believe that *p', whose truth is not guaranteed by the form of the judgement as with basic self-knowledge)?* How might the transcendental argument account for their groundless character? Again, seeking an answer to this question requires some reading between the lines. The argument for entitlement is at least consistent with groundlessness in the sense that it derives from a capacity to operate in accord with reason that one need not be capable of articulating. But this has more to do with the reasoning a subject might engage in about her first-order mental states than with judgements about in what that

first-order content might consist. One might argue that critical reason requires true judgements about that content, but this does not address the groundlessness of those judgements. An option here would seem to be the following. On Burge's view, our first-order mental states and non-contextually self-verifying judgements about them are ontologically distinct states that are causally linked together. Assuming the plausibility of this view (which, given his denial that such a causal link can play any part in an account of entitlement, is not immediately evident), Burge may be free to argue that the groundless character of our self-ascriptive utterances is explained by the underlying (or background enabling) presence of this causal mechanism. Such a mechanism, while not factoring in an account of epistemic entitlement, would vindicate the social epistemic practice of treating self-ascriptions as immediate or groundless. It would also account for the psychologically non-inferred character of self-ascriptive utterances.

It appears, then, that there is room in Burge's conception of the constitutive link between critical rationality and self-knowledge for an explanation of the key distinguishing features of self-ascriptive utterances. It also avoids one of Boghossian's main criticisms of his epistemologically "insubstantial" basic self-knowledge account. Recall Boghossian's claim that, however plausible Burge's account of basic self-knowledge might be, it could not have the kind of paradigmatic status Burge assigns to it. Contextually self-verifying judgements are by their nature infallible (or very nearly so – see Burge: 1998c, 240 for a minor qualification); but this is not in keeping with our ordinary understanding of the fallibility of most types of self-ascriptive claims. In the case of non-contextually self-verifying self-ascriptions there is a 'non-contingent'

warrant, but one that entails only that one must normally know what one's first-order states and the connections between them are. Burge therefore acknowledges that this need not always be the case, thus making room for mistakes in judgement connected to occasional failures of critical reason.

But this points to a questionable aspect of Burge's general approach to the issue. While he is surely correct in recognizing our limited fallibility with respect to self-knowledge, it seems questionable that the level of authority we enjoy with respect to our self-ascriptions can be correlated with the degree to which we reason critically. According to Burge, critical reason involves sophisticated cognitive abilities. For example,

> As a critical reasoner, one not only reasons, one recognizes one's reason as reasons. One evaluates, checks, weighs, criticizes, supplements one's reasons and reasoning. Clearly, this requires a second-order ability to think about thought contents or propositions and rational relations among them. ... When one engages in practical deliberation, one articulates and weighs considerations on each side, goes over possible sources of bias, thinks through consequences. Essential to carrying out critical reasoning is using one's knowledge of what constitutes good reasons to guide one's actual first-order reasoning. ...For reasoning to be critical, it must sometimes involve actual awareness and review of reasons; and such a reviewing standpoint must normally be available. (Burge: 1998c, 246-247)

Furthermore, as mentioned above, Burge argues that the ability to reason critically about one's first-order mental states is a necessary condition for the possibility of thinking first-

person present-tense thoughts about those states.[19] At least on the surface this seems to be a rather tenuous position to hold. Small children may, as Burge says, "reason blind," that is, non-critically (ibid., 247); but what are we to say about the status of their self-ascriptive utterances? Little Anna, who is unable to engage in the kind of reasoning found in Burge's description of critical rationality, says, "I don't want to wear pants. I want to wear a dress." This statement certainly appears to be true; it seems to accurately report her mental state at the time, and from it and her other behaviour it is reasonable to interpret her as desiring to wear a dress, not pants.

It would seem that Burge would have to deny this. At the very least, he would have to deny that such an utterance was an expression of a second-order judgement about her first-order desire. If so, he will need an explanation of why Anna's utterance, which on the surface resembles any other person's self-ascriptive utterance (for example, in its grammatical form, truth-evaluability, the security it exhibits, its immediacy, the role it plays in another's interpretation of her first-order state), is fundamentally different from a similar pronouncement by a mature critical reasoner. At the very least, such considerations should raise questions regarding the claim that critical rationality as described by Burge is needed for the authority and immediacy (groundlessness) of the social epistemic practice associated with self-ascriptions.

2.2.k Two More Concerns

[19] Incidentally, this suggests that critical reason is a pre-condition for the possibility of thinking not only non-contextually self-verifying thoughts, but contextually self-verifying ones as well.

There are two more issues that I will only briefly discuss here, as they will be explored in further detail once I have provided the argument for the expressivist account of first-person authority. The first concerns what at least appears to be the threat of an infinite regress in Burge's account. Burge argues that the warrant (entitlement) for, and reliable truth of, a subject's second-order beliefs and judgements about the rational standing of her first-order mental states is explained by the regulative or supervisory role they play in the maintenance of her rationality. To quote Burge:

> if one lacked entitlement to judgements about one's attitudes, there could be no norms of reason governing how one ought to check, weigh, overturn, confirm reasons or reasoning. For if one lacked entitlement to judgements about one's attitudes, one could not be subject to rational norms governing how one ought to alter those attitudes given that one had reflected on them. If reflection provided no reason-endorsed judgements about the attitudes, the rational connection between the attitudes reflected upon and reflection would be broken. So reasons could not apply to how the attitudes should be changed, suspended, or confirmed *on the basis of* reasoning depending on such reflection. But critical reasoning just is reasoning in which norms of reason apply to how attitudes should be affected partly on the basis of reasoning that derives from judgements about one's attitudes. So one must have an epistemic entitlement to one's judgements about one's attitudes. [Furthermore], if reflective judgements were not normally true, reflection could not add to the rational coherence or add a rational component to the reasonability of the whole

process. It could not rationally control and guide the attitudes being reflected upon.... (Ibid., 249-250)

On the one hand, the warrant for and security of self-knowledge are explained by the role it plays in critical rationality, which itself is required to regulate one's first-order mental life. The reflective second-order judgements we arrive at regarding our first-order mental states could only fulfill their regulative role if they counted as knowledge. That is, if the second-order judgement that one ought or ought not believe that *p* were not itself true and warranted, it could not serve as a reason to form, maintain, or discard the belief that *p*. So second-order judgements count as knowledge in virtue of the regulative role they play. But, we might ask, what explains the fact that our second-order judgements and reasoning about the first-order states we judge ourselves to have are so reliably correct? Burge argues that the rationality of a subject's first-order states is maintained by the supervisory function of second-order judgments. But this presupposes that those second-order judgements are themselves in accord with the norms of reason. But what explains this? That our first-order states are in accord with reason is explained by the supervisory activity of our second-order judgements. But what explains how those second-order judgements are normally sound? That they must be is dictated by the role they are said to play. But to make this point is not to explain how they remain so. If our second-order beliefs were not rational, they could not serve the regulatory role that Burge assigns them; but if their rationality is explained in the same way that the rationality of our first-order beliefs is explained, then a third-order of belief will be needed to regulate our second-order beliefs. But since our third-order beliefs can only perform that

regulatory role if they are themselves rational, Burge's account is set off on an infinite regress. Put another way, if our second-order judgements do not themselves require a third overseeing level that regulates and ensures their remaining in accord with reason, then they (somehow) remain in accord with reason without any higher-order supervision, and so such supervision cannot be necessary for rationality. So the appeal to higher-order supervisory intervention need not be required to account for the rationality of our first-order states.

This brings me to the second issue. An underlying assumption of this supervisory model is that our second-order judgements about our first-order states serve as reasons – in fact, as will be argued later, the primary reasons – for holding those states. At the end of reflection, the subject arrives at a judgement about what she ought to believe, desire, intend or feel, and this is what should directly motivate one's maintaining, adopting, or discarding a first-order state. However, as David Owens (2000) argues, it is questionable that such second-order judgement can do the controlling and guiding work that Burge and others who argue for the supervisory model of rationality assume it does. Especially with regard to the question of belief, it might be asked: if one's first-order reasons for believing that *p* were insufficient to motivate a belief that *p*, why would the second-order recognition that one has such reasons do any better? As Owens remarks, such second-order judgement looks to be an idle wheel in our motivational economy, and as such is one we can do without (Owens, 18). As mentioned, these matters will be explored further in the final chapter, where the supervisory model of rationality will be discussed in light of the expressivist reading of self-ascriptions. But if, as these points suggest, the

supervisory model of rationality is in trouble, then so are the purported ties between rationality and self-knowledge that make use of that model of rationality. And without these ties between rationality and self-knowledge, we are (so far) left with no good explanation of the immediacy and authority of self-ascriptions.

2.3 Bilgrami: Self-knowledge and the Grammar of Responsible Agency

According to Akeel Bilgrami (1999), a proper account of the special character of self-knowledge requires understanding the essential role it plays in what he takes to be a Strawsonian conception of responsible agency. On this view, responsible agency is a thoroughly normative idea; rather than basing it on some metaphysical property of persons to which we appeal to justify our practices of punishment and blame, we ought to see our conception of agency as derivative of those practices themselves. Given its status as a necessary condition for those practices, it is argued that self-knowledge is itself conceptually grounded in normative considerations tied to responsible agency. This shows self-knowledge to play a definitional and constitutive role in the intentional states self-known – there is a necessary conceptual link between our status as responsible agents and self-knowers that can be encapsulated in the following bi-conditional: under conditions of responsible agency, one believes that one believes that *p if and only if* one believes that *p*. Similarly, if one believes that *p*, then one believes that one believes it. What constitutes such true beliefs as knowledge is the necessary connection they bear to our status as responsible agents.

2.3.a Strawson's Normative Reconciliation of Freedom and Determinism

Bilgrami sees his analysis of self-knowledge as an extension of Strawson's analysis of responsible agency (Strawson, 1959). He argues that our understanding of the asymmetrical character of self-knowledge ought to be placed in the same conceptual framework as Strawson's discussion of the idea of non-coercive causality and how it relates to our understanding of agency. For the asymmetries thought to characterise self-knowledge can be explained by the role self-knowledge plays in a normative conception of responsible agency (Bilgrami, 215 ff).

According to Bilgrami, Strawson argues for a reversal in the order of explanation of our idea of responsibility that underlies our conception of agency (ibid., 214). This reversal is proposed as a way to reconcile the apparently opposing ideas of freedom and determinism. How do we justify the practices of, for example, punishment and reward associated with the assignment of freedom and responsibility, with the idea that all actions are subject to universal causal laws? One compatibilist strategy is to admit universal causality, but then argue for a metaphysical distinction between coercive and non-coercive causes. It could then be argued that one could justifiably be held responsible for one's actions where they were the result of non-coercive causality. However, according to (Bilgrami's) Strawson, this distinction will not work. As Bilgrami puts it,

> there is nothing in what are identified as non-coercive causes which by itself makes evident that something like punishment is just for the sorts of harmful actions that

> we tend to punish. ... [N]othing that we can find just by staring at the causes justifies this distinction between causes. The so-called non-coercive causes do not wear the relevance of their non-coerciveness on their sleeves in a way that justifies our practices surrounding the assignment of responsibility. (ibid., 213-214)

When looking at what is identified as a non-coercive cause, there is nothing on the face of it – that is, examined independently of the evaluation of the action as one for which one should or should not be held responsible – that marks it as non-coercive. And nothing about the non-coerciveness per se (that is, conceived purely as a metaphysical property) justifies our normative practices surrounding the assignment of responsibility (for example, punishment or blame). Bilgrami contends that arguing otherwise amounts to committing a naturalistic fallacy (ibid., 215).

One might seek to avoid this problem by arguing for a conception of responsibility that can stand independently of such normative considerations. However, Bilgrami contends that such a disconnection is untenable,

> [f]or we must ask what interest there is in such a stipulated notion of responsibility which bears no relation to our practices at all. I think that there cannot be any answer to this question which does not point to *some* normative significance that that revised notion of responsibility has for us. (ibid., 216)

In light of these problems it is recommended that, instead of attempting to reconcile freedom and determinism by invoking a suspect metaphysical distinction between non-coercive and coercive causes, the compatibilist should relinquish the idea that our conception of ourselves as responsible agents has such a metaphysical basis. We ought to

give up on non-coercive causality and freedom of the will as metaphysical ideas and instead see them as derivative of the normative considerations to which we appeal to justify the practices associated with the assignment of responsibility. We should recognise that it is these practices, and the justifications that we offer for them, that ought to ground and inform our notion of freedom and responsible agency. On Strawson's recommendation, there is no understanding the idea of, say, non-coercive causality as a necessary condition for free action apart from consideration of the practices of praise and blame it is invoked to justify. The very idea of a non-coercive cause is grounded in the normative evaluative attitudes, and any account of it must, at bottom, make reference to this normative dimension/foundation.

2.3.b Bilgrami's appropriation of Strawson – Self-Knowledge and Responsible Agency

Bilgrami argues that we should think of self-knowledge in a similar light. As with the analysis of the idea of non-coercive causality, a proper understanding of the nature of self-knowledge ought to take into consideration the "subsidiary" role it plays in a normative conception of responsible agency (ibid., 215). This has a dual benefit, in that (i) it avoids the sort of problem outlined above regarding how self-knowledge can serve as a justification for certain normative practices, and (ii) provides an account of the authoritative and immediate or groundless character of self-knowledge.

The analysis begins by focusing on self-knowledge as a necessary condition for holding one another (and ourselves) responsible for our actions. The initial grounding

claim is that "[s]elf-knowledge is necessary for responsibility *for no other reason* ... than that our *evaluative* justifications of the practices of assigning punishment and blame seem to be apt only when self-knowledge is present" (ibid.). Recall that, as Bilgrami sees it (following Strawson, as he reads him), responsibility and freedom are normative ideas; they cannot be understood apart from the evaluative reactive attitudes and practices associated with them (for example, of blame and punishment). It is these "reactive attitudes of evaluation of people (including ourselves) and their (our) action ... [that] underlie and justify both the general ascribability of responsibility and freedom to our actions and the practices of punishment surrounding them" (ibid., 214). According to this normative conception, the idea of responsible agency ought to be seen as derivative of these evaluative reactive attitudes and the justification we offer for them (since a metaphysical explanation of responsibility and freedom cannot account for the evaluative practices associated with them, at least not without involving a naturalistic fallacy).

The practices of assigning blame and punishment for a person's (including one's own) actions or conclusions (and various other sorts of evaluative attitudes and actions that define a normative conception of responsibility) are grounded in our reactive attitudes and the internal justifications we offer for them. Those reactive attitudes and their justifications are only deemed appropriate when the subject whose action or conclusion is the object of the reactive attitude has knowledge of the intentional states that explain or rationalise it (ibid.,215, 221). In other words, the internally justifiable reactive attitudes that underlie the practices that define a normative conception of responsible agency are themselves justified only where self-knowledge is present.

Bilgrami claims that this shows that the idea of self-knowledge has a fundamentally normative character – there is a necessary conceptual relation, based on normative considerations, between self-knowledge and the intentional states that lead to actions (or conclusions) involved in responsible agency.

We find, then, that self-knowledge is constitutive of a certain class of intentional state, namely those involved in responsible agency. Under conditions of responsible agency – that is, whenever an intentional state potentially leads to an action or conclusion that can be the object of internally justified reactive attitudes – there must be self-knowledge (true second-order belief that is in some sense justified or warranted) of those intentional states. So, how does this necessary conceptual relation inform our understanding of self-knowledge? For one thing, it rules out what Bilgrami calls a spare causal-perceptual type of explanation (ibid., 209-210, 227-228). Such a model need not include any sort of empirical cognitive achievement analogous to perception; it is enough that the main explanatory work be done by a causal mechanism that links lower-level intentional states to higher-level beliefs about them. On the causal-perceptual view, the formation and security of second order beliefs, and what makes them count as knowledge, are explained by the reliable functioning of such a second-order belief-forming causal mechanism. For Bilgrami, the key problem is that, as with any causal mechanism, there must be the possibility of failure. That is, it must be possible for one to have a lower-order intentional state that, due to malfunction in the higher-order belief-forming mechanism, does not become the object of a higher-order belief. Given this possibility of breakdown, there must be a certain independence of second-order beliefs

from the first-order intentional states they are about that bars Bilgrami's constitutive claim.

That said, it should be mentioned that the necessary relation between self-knowledge and responsible agency normatively conceived does allow for what Bilgrami terms "a frankly acknowledged fallibility" (ibid., 226). That is, it allows for the possibility of intentional states of which the subject may be ignorant – not all first-order intentional states are necessarily self-known. In fact, it is just those unknown states that figure in self-deception (and for which responsibility lapses). According to Bilgrami, self-deception is a form of self-blindness; it consists not in having false second-order beliefs, but rather results from a lack of self-knowledge regarding intentional states one has that conflict with those that are self-known (ibid.218). This will be discussed in greater detail below.

What, then, makes self-knowledge *knowledge*? As Bilgrami notes, something more than the mere presence of a true second-order belief is needed for that belief to count as self-knowledge (ibid., 224). What further thing (as he puts it) turns a true second-order belief into knowledge (ibid.)? In a causal-perceptual account it would be that the second-order judgement resulted from the appropriate causal relation linking the first- and second-order states. On his constitutive view it is the role the second-order belief, as knowledge, plays in our conception and practices of responsible agency. For "[i]t is this condition of responsibility being fulfilled ... which presupposes that there is self-knowledge" (ibid.). But this seems somewhat circular. On the one hand, it is argued that self-knowledge is a necessary condition for responsible agency. One can only be

interpreted as an agent and held responsible for one's actions or conclusions when they derive from intentional states that are self-known. On the other hand, the knowledgeable status of those true second-order beliefs depends upon the role they play in the justification of the very thing they are said (as knowledge) to justify, namely the internally justified reactive attitudes that ground a normative conception of responsible agency. The internally justified reactive attitudes are only apt when self-knowledge is present; but what makes second-order belief knowledge in this case – what plays the part of warrant here – is the role it plays in the justification of the internally justified reactive attitudes. How can that for which warranted true second-order belief is a requirement (viz., responsible agency) play a role in the constitution of that self-knowledge?

Put another way, don't we need an independent, prior determination of the presence of self-knowledge to determine that the internally justified reactive attitudes are appropriate, that a subject may be taken as a responsible agent? And if so, might not a spare causal-perceptual account provide a tidy explanation of how we get that self-knowledge after all? Where the causal mechanism functioned properly, self-knowledge, and thus responsibility, would be in place. And in the event of a breakdown in the mechanism, self-knowledge would fail and responsibility lapse (as in self-deception as he sees it).

Bilgrami acknowledges the "temptation" to argue this way, but contends that the temptation only arises if one forgets about what he calls the subsidiary status of self-knowledge as a necessary condition for responsible agency (ibid., 229). To repeat, the idea of self-knowledge he proposes originates in considerations that are essentially

connected to a normative conception of responsible agency. As indicated above, on this picture by definition there can be no breakdown in the relation between the second-order beliefs and the first-order intentional states that are their objects. As he says, under conditions of responsible agency "they stand or fall together" (ibid., 231). This feature, which is an essential aspect of self-knowledge normatively understood (that is, in terms of its place in a normative conception of responsible agency), cannot be accommodated in a causal-perceptual model of explanation. That is why there is a "failure of fit" between the causal account and self-knowledge so conceived; the inductive reliabilism of the former (what, on that model, makes a true second-order belief knowledge) does not match the necessity of the connection between first-and second-order states that the latter demands.

Say we grant this. What, then, about the question of circularity? Again, Bilgrami argues that this only appears problematic if we take our eye off of the "subsidiary" nature of self-knowledge as a necessary condition for responsible agency. The status of self-knowledge as a necessary condition is grounded on evaluative considerations tied to a normative conception of agency. This necessary relation between self-knowledge and the internally justified reactive attitudes defines the kind of intentional states involved in self-knowledge. To quote Bilgrami, on the view recommended "there is no way to understand the attribution of higher-order beliefs in self-knowledge without understanding that they take as their embedded objects the kind of intentional states essentially caught up with the reactive attitudes" (ibid., 225). On this picture, self-knowledge just is knowledge of one's intentional states involved in responsible agency. Thus, if a true second-order belief is to

count as self-knowledge, the condition of agency must be in place. And, Bilgrami contends, that self-knowledge is itself required for that condition to obtain does not affect this necessary conceptual relation. So, it may in a sense be circular, but it is not viciously so. For, as will be discussed below, it turns out that while self-knowledge is needed for agency, agency is also needed for self-knowledge.

2.3.c Agency, Intentionality and Self-Knowledge

To this point we have seen it argued that, under conditions of responsible agency, self-knowledge is constitutive of intentionality. There is a kind of intentional state – namely those involved in agency – which must be self-known. But, in a sense, Bilgrami's claim is broader than this; because intentionality presupposes responsible agency, self-knowledge is constitutive of intentional states in general. In support of this Bilgrami asks us to imagine a subject who lacks the agent's point of view (ibid, 235 ff.). He is only able to take a third-person perspective on himself, so for him all of his thoughts and actions are mere objects that happen in him. His thinking is wholly passive, and he lacks any sense that he can make a difference in his future. As imagined, thoughts occur to him, but he is unable to "actively" assess them, that is, evaluate them in light of self-reactive attitudes he has toward them. Suppose such a subject spoke and acted in a rational manner, and even made avowals regarding his intentional states – why not say that he has self-knowledge, and offer a causal-perceptual account to explain it? If this were possible, the general constitutive thesis would be false. What rules this out?

Bilgrami contends that the attempt to imagine such a wholly passive subject defeats us; such a subject would not count as having genuine thoughts. But this is not simply due to the foreignness of such supposed subjectivity. Rather, an analysis of self-ascriptions and second-order belief reveals that first-order intentional states themselves, and not just actions or conclusions that may derive from them, are potential objects of the internally justified reactive attitudes. To begin, he suggests that "one would not know what role second-order beliefs could have in our psychological economy if they did not emerge in actions that indicated the existence of their embedded beliefs – or, if not in actions, then in a *preparedness* to act on the beliefs (or desires) that are embedded" (ibid., 230). The term 'preparedness' is key – by it he means that one would either be disposed to act on the first-order state, or, lacking this, be prepared to accept criticism for lacking that disposition and make some effort to cultivate it. Given this role for second-order belief, it follows that a sincere avowal, expressing a genuine second-order belief, would have to reflect that one was prepared (in the sense just outlined) to deploy the first-order intentional state in one's thinking and actions. Where one was not so disposed, rather than suppose that a subject had avowed a false second-order belief, one would be inclined to suppose the avowal insincere and thus as not expressing a second-order belief.

How does this view of avowals and second-order belief support the view that agency is a necessary condition for intentionality? The key term is 'preparedness'. The idea that a second-order belief is attributable so long as the subject is prepared to accept criticism for not having the disposition to act on the first-order belief avowed "reveals all the real depth of the normativity of intentional states. This very natural intuition concerning

avowals implies about first-order intentional states that we can fully possess an intentional state *as a commitment*, as a normative stance or stances..." (ibid., 231, italics added). In other words, it shows that first-order intentional states themselves can be the object of internally justified reactive attitudes, and that "thought and intentional states are caught up directly in agency just as much as actions are" (ibid., 239).

Suppose we admit this claim about the fundamentally normative character of intentional states. Does it rule out the possibility of false sincere avowal as suggested? Take once more the issue of self-deception. Suppose that someone sincerely (by all appearances) avows that he loves and respects his spouse, but all of his actions indicate otherwise ("a conically crude example," as Bilgrami might say [ibid., 218]). We describe him as self-deceived. But this does not mean we should say that he has false second-order beliefs regarding his first-order states. Instead, in accord with what Bilgrami takes to be a natural intuition regarding the infallibility of second-order belief, we ought to say that he does have the states he reports himself to have, but he also has those other incompatible states to which he is self-blind. Of course, this assumes that he is prepared to act on the first-order states self-ascribed in the sense mentioned above.[20] Supposedly this would consist in a willingness to cultivate those dispositions that reflected the first-order states self-ascribed and to which he says he is committed. But how do we determine that those

[20] Bilgrami is less than clear on this point. In his initial, unqualified (in the sense that the agency condition is not yet in play) discussion of self-deception, he claims that the self-ascription is true, in spite of all the evidence to the contrary. In other words, at this point the only evidence for the second-order belief seems to be the subject's avowal of it. But suppose this is qualified, so that the "preparedness condition" must also be met. The inconsistency in words and actions is pointed out to the subject, and she accepts the criticism and strives to cultivate the appropriate dispositions (those inconsistent with her previous actions). At this point the self-blindness and self-deception would end – presumably, in accepting the criticism, she would come to recognise those states underlying her actions that were inconsistent with her avowals.

actions reflect the presence of the first-order state (of love and respect), rather than a commitment to the truth of the second-order belief itself? Why might they not be evidence of wishful thinking expressed by a false avowal that the subject nonetheless takes to be true?

The only thing that seems to bar this is a claim about the role second-order belief plays in our conceptual economy. But could we not admit this general claim and at the same time allow for this kind of false belief? Could we not admit that we would not know what role second-order belief would play in our conceptual economy if they were not in the main true (did not emerge in actions that indicated their truth), but that they also allow for instances of false belief as well (false second-order belief that played its own role in our psychological life, as in the scenario just imagined).[21] Nothing in Bilgrami's general argument regarding the necessary connection between self-knowledge and agency – that any state that can be the object of internally justified reactive attitudes must be self-known – seems to rule this out. One might argue that, assuming it does not negatively affect his general thesis, this is not such a bad thing for Bilgrami's view. For, as Boghossian points out, most would simply reject the idea that we are infallible with respect to our self-knowledge claims (Boghossian, 151).

2.3.d The Infallibility of Second-Order Belief – A Potential Difficulty

[21] In other words, why not import Davidson's anti-sceptical argument for the general veracity of first-order belief into this higher-level of judgement (see Davidson: 2001a, 2001h)?

Bilgrami argues that the reliability of and warrant for second-order belief is explained by the role self-knowledge plays in a normative conception of responsible agency. When conceived along these lines, we find that self-knowledge is constitutive of the states it is knowledge of – it is part and parcel of the idea of an intentional state involved in agency that it be self-known. Bilgrami summarises this constitutive thesis as follows: under conditions of responsible agency, (1) if a subject believes that p, then she believes that she believes it, and (2) if she believes that she believes that p, then she believes it. As it stands, this is a thesis about a certain class of intentional states. To make it fully general (and thus rule out the possibility of a causal perceptual account), he argues that agency is a necessary condition for intentionality. However, this is not to say that every intentional state must by definition be self-known; as the discussion of self-deception shows, self-unknown intentional states are possible. This, Bilgrami argues, allows for a "frankly acknowledged fallibility" with respect to self-knowledge (ibid., 226). But, as noted above, in denying the infallibility of self-knowledge he only rejects the idea that all first-order intentional states are necessarily self-known. It does not allow for false self-knowledge claims.

This, I have suggested, is open to doubt, as it is supported by a questionable claim regarding the role of second-order belief in our mental lives. If the doubt is warranted, it will negatively affect the argument in support of his "radical assumption about the deep relation between agency and thought" (ibid., 239), for this seems to require the infallibility of second-order belief. How so? Bilgrami contends that sincere avowals expressive of second-order belief are attributable just in cases where (1) the subject acts,

or (2) is prepared to act, in ways that indicate the existence of the first-order intentional state self-ascribed. This "very natural intuition concerning avowals," (ibid., 231) is said to reveal the deeply normative character of intentional states. However, if false second-order beliefs are possible, then it becomes problematic how one might distinguish actions indicative of a first-order intentional state-as-commitment from those that may reflect a subject's commitment to the truth of what is in fact a false second-order belief. Bilgrami argues that the fact that the subject is prepared to act in response to the criticism that he is not acting in accord with his avowal is evidence that he has the first-order state avowed. The question is: why couldn't his subsequent preparedness to act according to his avowal (in other words, subsequent to the criticism) simply be evidence that he is self-deceived, or has succumbed to wishful thinking? He avows that his wife loves him, but his other actions and behaviour indicate otherwise. This is pointed out to him, upon which he seeks to change his behaviour to match his avowal. But this may be because he wants to believe that his wife is faithful, so avows it and acts accordingly, even though he really believes her a cheat. In other words, a question arises about whether a subject's preparedness to act a certain way in response to criticism over his failure to act in accord with his avowal would reflect (i) the presence of the first-order state self-ascribed rather than (ii) the subject's (false) second-order belief that he has that state. A lack of clear individuating criteria for the two sorts of actions places the plausibility of first-order states-as commitments in doubt.

Say this problem could be overcome. Perhaps another argument in support of the intentionality-as-commitment claim could be made; or, perhaps the *reductio* cited above

might suffice to show the necessity of agency for intentionality. While one half of Bilgrami's biconditional would have to be abandoned, the basic claim regarding the relation between self-knowledge and agency would be left intact. While the possibility of false avowal and second-order belief would be admitted, one would still need to be interpretable as an agent in order to have self-knowledge. The necessary conceptual relation between self-knowledge and agency could still serve as the proper framework from which to understand self-knowledge and the social epistemic practice associated with it. That is, it would serve as a vindicating explanation of why we take one another's self-ascriptions as generally (if not absolutely) authoritative and groundless; our treating one another as rational agents would require it.

But now it may seem that we have moved from a view that over-described first-person authority to one that arguably under-describes it. Once again, as with Burge, we find that the reliable truth of avowals is explained by a capacity (in this case, the capacity to take on a sophisticated first-person perspective on their mental states) that many, in spite of their ability (by all appearances) to avow intentional states, may lack. According to Bilgrami's general view of avowals, a sincere utterance of "I want ice cream" expresses a second-order belief about the speaker's first-order desire – so long, that is, as the speaker is interpretable as an agent. Take now a child, who, while not yet capable of taking the first-person perspective as Bilgrami describes it, says "I want ice cream." She is given the ice cream, and walks away with a smile on her face. What ought we to say about this behaviour? According to Bilgrami she lacks the intentional states usually expressed or reported by an adult's use of such language (or, as Bilgrami says, one who is

"more or less" an adult, namely is capable of self-reflective or critical rationality [ibid., 236]). If so, then she ought not be interpreted as speaking a language at all. For, as Bilgrami writes, a speaker will not be interpreted as meaning various things "unless we interpret her as having various complicated sets of interrelated beliefs and desires" (ibid., 229). It therefore becomes a mystery why we should take the child as meaning anything by her use of words, or as intending to accomplish anything by them, let alone suppose that she got her intentional states right. At any rate, some other story will need to be provided to account for this proto-linguistic behaviour.

2.4 Conclusion

In this Chapter I have examined the various arguments Shoemaker, Burge, and Bilgrami offer in support of the general idea that the unique features of self-knowledge are explained by the role it plays in critical rationality and rational agency. While I have not yet addressed the general plausibility of this connection (this will be undertaken in Chapter 4), I have suggested that in each case there are significant difficulties with their arguments that should place the link in doubt. At the very least, they suggest that other possibilities need to be explored.

In the chapter to follow I first look at another account – that of Richard Moran – that also argues for a strong connection between self-knowledge and rationality. However, Moran's understanding of this connection takes a much different form. Instead of focusing on the supposed supervisory function of second-order belief, he argues that we should see our authoritative self-ascriptions as expressing a unique sort of higher-order

rational commitment with respect to our mental states, namely that they be determined by our understanding of the first-order reasons for them. In certain respects Moran's shift away from the supervisory model and emphasis on first-order deliberation constitutes an advance in the discussion. However, I shall argue, it suffers from problems that are in large part attributable to the perceived need for an *epistemic* account of authoritative self-ascription. I therefore turn to the consideration of Donald Davidson's non-epistemic account of self-knowledge, which, though incomplete, does point the way to the explanation of our capacity for authoritative self-ascription.

Chapter 3: Toward an Account of Authoritative Self-Ascription – Moran and Davidson

Introduction

In this chapter I look at Richard Moran's and Donald Davidson's accounts of self-knowledge. I begin with Moran, whom I described earlier as a transitional figure in my overall discussion of the supposed relation between self-knowledge and rationality. Like Shoemaker et al., he argues for intrinsic ties between self-knowledge and rational agency. However, his understanding of these connections differs: rather than focus on the supervisory role self-knowledge plays in the maintenance of rationality, he focuses on the role commitment to first-order reasons plays in the formation of second-order beliefs about the content of our own minds. As he sees it, past treatments of self-knowledge – both epistemic and deflationary – have failed to appreciate the importance of this connection. He argues that this was at least partly due to certain Cartesian assumptions about the mental that linger even when the perceptual or inner sense model of introspection is explicitly rejected. Moran's view of the relation between self-knowledge and rationality leads him to what I call a "double expressive" account of authoritative self-ascription, where utterances of 'I believe (desire, intend, etc.) that *p*' express both the first-order state ascribed as well as the second-order belief that one has that state. I contend that, in a sense, he is half-right. I agree that we need to focus on first-order reasoning to understand authoritative self-ascription; however, I suggest that his insistence on the presence of an additional higher-order level of belief is itself a vestige of Cartesianism that we may do without.

To see how, I turn to a discussion of Davidson's deflationary account of self-knowledge. Davidson also argues that a latent Cartesianism continues to inform discussions of self-knowledge. He argues that a successful explanation of authoritative self-ascription must overcome the Cartesian view of mental states as objects before the mind. His appeal to the semantic authority a speaker enjoys with respect to her own use of words in the explanation of first-person authority does just that. However, as mentioned previously, I argue that his account still suffers from a particular defect that a properly construed expressivist understanding of authoritative self-ascriptions resolves. This version of expressivism is taken up in the first part of Chapter 4.

3.1 Moran: Self-Knowledge and the First-Person Perspective

3.1.a The Possibility of Non-Cartesian Introspection

Like Shoemaker, Burge, and Bilgrami, Richard Moran (2001) argues that a proper account of self-knowledge must take into consideration the part it plays in our understanding of rational agency. However, Moran concentrates his analysis on the cognitive means by which we arrive at self-knowledge, and how that is essential to our conception of rational agency and the first-person perspective. In short, Moran claims that self-knowledge derives from a substantial cognitive achievement definitive of the first-person perspective and rational agency.

Understanding the nature of this cognitive achievement requires broadening the scope of the inquiry from purely epistemological issues to the nature of the first-person

perspective more generally. This requires that we move beyond what he calls the "Cartesian and empiricist legacy" that he thinks continues to inform many philosophers' thinking about self-knowledge (Moran, 3). The problem here is not that philosophers continue to adhere to the Cartesian picture of inner perception or the infallibility of such judgments – as he notes, these ideas have been subject to serious criticism and rejected by most philosophers. Rather, the problem is the close association, if not identification, many philosophers have made between introspection and the Cartesian/perceptual understanding of it – that if introspective judgement is to be anything at all, it must conform to some variant of this basic perceptual model. Thus, with the demise of the perceptual model has come the trend toward deflationary accounts that see the phenomena normally associated with self-knowledge (immediacy, groundlessness, and authority) as deriving from something other than the subject's privileged access to her own mental states (ibid.).

For Moran this is a mistaken move – we ought not throw out the introspective baby with the Cartesian/perceptual bathwater. Traditionally introspective access has been conceived along the lines of an ill-fitting third-personal/empirical model of explanation imported into a mental interior. While modifications were made to accommodate the immediacy and authority thought to be characteristic of such judgements, they still had nothing to say about what he argues are other equally significant first-personal considerations regarding the connection such judgements have to the status of the subject as rational agent. And the same can be said for deflationary accounts – neither have they had anything to say about how the capacity to authoritatively self-ascribe one's mental

states is essential to these aspects of the first-person perspective. However, if we expand the analysis to include these features (those having to do with the subject as rational agent), we will find that what are normally called self-knowledge claims do involve non-perceptual, substantial judgement (that is, judgement that involves genuine cognitive achievement) unique to the first-person perspective. As we shall see, for Moran this means understanding the role our reflective second-order judgements on what first-order states we ought to have play in arriving at our self-ascriptions. Moran's approach thus contrasts with those of Shoemaker, Burge, and Bilgrami, who argue for a necessary conceptual relation between self-knowledge and rationality that makes no appeal to any sort of detective work in the explanation of authoritative self-ascription of one's first-order state.

3.1.b. Moran on Insubstantial Approaches to Self-Knowledge – Boghossian and Burge

We can get an initial sense of how Moran conceives of the substantiality of self-knowledge by looking at criticisms he offers of a couple of deflationary approaches. He first examines Boghossian's discussion of "insubstantial" self-knowledge, or self-knowledge that is "based on nothing" or "nothing empirical" (that is, neither on observation nor inference from observation). This includes such "indexically grounded" judgements as 'I am here now' and Burge's "basic" self-knowledge (those judgements, such as 'I am thinking that *p*', that are contextually self-verifying – that, simply in virtue of being thought, are made true). Rather than being based on some sort of "awareness of

some independently obtaining state of affairs" such judgements "share the feature that the appearance of knowledge is grounded purely logically (or transcendentally), ... [where] the denial of any such statement would involve some kind of immediate incoherence" (ibid., 17).[22] Moran notes that if this is what makes such judgements "insubstantial," then it follows that, by contrast, one feature a substantial first-person self-knowledge claim ought to have is that "its truth conditions be in some way independent of the making of the judgement" (ibid., 18). This, he maintains, "is a form of cognitivity that any account of introspection as a source of knowledge would seek to preserve" (ibid.).

This is a key claim for Moran, and the way in which he construes it will be important for his account of the substantiality of self-knowledge elaborated below. For now, we can note that by "independence of truth conditions" here he means that the knowledgeable status of the judgement does not derive from its form – that merely making the judgement does not guarantee its truth or justification (as is said to be the case with the type of statements just considered). For in another sense the truth conditions of such "insubstantial" judgements *are* independent of the making of the judgement. For example, 'I am here now' is true if and only if I am here now. But I need not make that judgement to be here now. Similarly, the (supposed) second-order belief expressed by 'I am thinking that writing requires concentration' is true if and only if I am thinking that

[22] Moran writes here of "the appearance of knowledge," which suggests that he thinks such utterances should not count as knowledge claims. Either that, or one could presume that by "knowledge" here he means "substantial" knowledge – that is, knowledge grounded on inference or observation. At any rate, although their accounts are "insubstantial," both Boghossian and Burge think that such judgements remain instances of genuine self-knowledge, namely true second-order belief that is in some sense justified.

writing requires concentration. But I may think 'Writing requires concentration' without making the second-order judgement that I do so think.

3.1.c. Moran on Wright's Deflationary Approach to Self-Knowledge

On Crispin Wright's constitutive account of self-knowledge, many of our authoritative self-ascriptions are insubstantial in the second sense just outlined (see, e.g., Wright: 2001d, 2001e). That is, a subject's first-order mental states are sometimes conceptually dependent for their existence and identity upon her second-order judgements about them that get expressed in her self-ascriptions – in certain cases, sincerely self-ascribing a state constitutes one as being in it. And this, Wright thinks, does lead to the conclusion that what is normally called self-knowledge is not grounded on any cognitive achievement, and so neither needs nor admits of any sort of epistemic explanation.

Wright is worth a detailed examination in this context for two reasons: First, as we shall see, Wright provides a good example of a radically deflationary approach to the purported "epistemology" of self-knowledge that Moran takes to challenge a "robust" realism of mental states, and so best illustrates the kind of position that Moran sets himself against. Secondly, Moran's own conception of what a more substantial epistemology of self-knowledge should be takes shape in large part against the backdrop of his reactions to Wright's deflationary account. (Wright is, in a sense, the best instantiation of his philosophical opponent.)

As Moran notes, Wright reaches this deflationary conclusion after considering various Wittgensteinian objections to a Cartesian, inner-perceptual understanding of authoritative introspection (Moran, 23). After finding this model of explanation for first-person authority wanting, Wright maintains that, "we require a different explanation, dissociated from introspection" (Wright: 2001a, 137). And, as far as he can see,

> There is only one possible broad direction for such an explanation to take. The authority that our self-ascriptions of meaning, intention, and decision assume is not based on any kind of cognitive advantage, expertise, or achievement. Rather, it is, as it were, a concession, unofficially granted to anyone whom one takes as a rational subject. It is, so to speak, such a subject's right to declare what he intends, what he intended and what satisfies his intentions; and his possession of this right consists in the conferral upon such declarations, other things being equal, of a *constitutive*, rather than descriptive role. (Ibid., 137-138)

Wright offers two sets of considerations in support of this view. One concerns the requirements for the interpretation of one another's intentional states. He begins by noting that "the *telos* ... of the practice of ascribing intentional states to oneself and others is mutual understanding" (Wright: 2001c, 313). Those ascriptions answer for their content to the behaviour (verbal and otherwise) that expresses the states they ascribe – there is a constitutive relation between the identity of a given intentional state and how a subject behaves (what he calls the "theoreticity" of intentional state concepts (Wright: 2001d, 340]). It is argued that interpretive practice requires that we allow each other's self-ascriptions to stand by default (what elsewhere he calls a "social concession"

[Wright: 2001a, 138]). This is because a subject's avowals (as Wright refers to what I am calling self-ascriptions) are indispensable for arriving at an interpretation of her intentional state.[23] As Wright says,

> taking the apparent self-conceptions of others seriously, in the sense involved in crediting their apparent beliefs about their intentional states, as expressed in their avowals, with authority, almost always tends to result in an overall picture of their psychology which is more illuminating – as it happens, enormously more illuminating – than anything which might be gleaned by respecting all the data except the subject's self-testimony. (Ibid.)

From this we are led to conclude that avowals are partly constitutive of what it is to be in a mental state. For without them we can make no sense of what it would be to be in an intentional state. There is no way to identify or individuate intentional states without a subject's self-ascriptions – they are an essential piece of the puzzle that an interpreter cannot do without. Thus, it makes no sense to say that self-ascriptions are extension-reflecting, that they express judgements about some independently existing first-order state of affairs. In the language game of intentional states, self-ascriptions are extension-determining – they play an essential role in determining what it is to be in a given intentional state. And one would only reject a subject's avowal if there were a positive reason to reject it, if (i) accepting an avowal at face value stood in the way of the

[23] It should be noted that Wright uses 'avowal' to refer to what I am calling authoritative self-ascription. Moran will use the term (sometimes qualified as 'genuine' avowals) to refer to those self-ascriptions that obey what he calls the Transparency Condition (see 3.1e below).

formulation of a coherent interpretation/rationalisation of her behaviour and (ii) a better understanding could be gleaned by rejecting what she said regarding her own state.

The second set of considerations – the one Moran focuses on in his critique – regards the *a priori* status of the provisional biconditionals that inform our notion of first-person authority. Wright begins by noting that when it comes to our practice of ascribing intentional states (self- and other-), there is a limited number of circumstances under which we might discount a subject's avowal (ibid., 139). Or, to put it in a positive way, there is a set of conditions – what he calls "(C) conditions" (for example, that the subject is not self-deceived, has a mastery of the relevant concepts, and is adequately attentive to the matter) – that when satisfied make it reasonable that a subject's avowals should be presumed correct (Wright: 2001b, 200-203). So we have the following conditional that defines our notion of first-person authority:

If C (Jones), then (Jones intends to P *iff* Jones believes he intends to P).

Wright then argues that the requisite (C) conditions are "positive-presumptive"; as he puts it, "such is the 'grammar' of ascriptions of intention [and other mental states], one is entitled to assume that a subject is not materially self-deceived, or unmotivatedly similarly affected, unless one possesses determinate evidence to the contrary" (ibid., 202). The positive-presumptive character of the (C) conditions makes it *a priori* reasonable to assume that they are met (assuming a lack of any evidence to the contrary); this, in turn, makes it *a priori* reasonable to hold the embedded provisional biconditional – 'Jones intends to P if and only if Jones believes he intends to P' – to be in effect.

Now, he goes on, one might read the biconditional in one of two ways. On the extension-reflecting reading, the left hand side describes a determinate state of affairs, which, if the (C) conditions are in place, the subject is able to detect. On this view, the (C) conditions collectively determine the conditions for a cognitive success, which an avowal may serve to report.

However, it might be read another, extension-determining, way:

> The alternative ... is to accord priority to the right-hand side. The resulting view would see the disposition to make the avowal as *constituting* the state of affairs reported by the left-hand side when the provisos are met. So the subject's cognition of an independent state of affairs does not come into the picture. Rather, he is *moved* to make the avowal and, subject to the provisos, it stands. (Wright: 2001a, 140)

The question is, which reading is compatible with the *a priori* status of the biconditional? As indicated above, Wright argues in favour of the latter reading. If the concept of intention works in such a way that, under the restricted set of (C) conditions, a subject's second-order judgement expressed by her avowal determines the content of her first-order state, then we have a neat explanation of why it is *a priori* reasonable to hold the biconditional true. Compare this to the supposition that avowals are extension-reflecting. In this case, what a subject intends will be determined independently of what she believes about her intention. But if this is so, he asks, what reason might we have to suppose it *a priori* reasonable to hold that a subject's belief about her intention must be true? Wright does not deny that a case for the extension-reflecting view that coheres with the *a priori*

status of the biconditional might be made; however, he maintains that, in light of the above considerations, the onus is on one who prefers that view to provide an account that (1) avoids the problems associated with Cartesian introspection and (2) explains just what determines the identity of the first-order states in question, if not those second-order judgements themselves (Wright: 2001b, 205).

Moran offers three criticisms of this view. First, and perhaps most fundamentally for him, he contests Wright's deflationary conclusion. Second, he argues that even if intentional state concepts are judgement-dependent in the way Wright suggests, this does not explain the asymmetries thought to obtain between first- and third-person ascriptions. Finally, he contends that the constitutive account does not explain why first-person authority is a "rational demand".

With regard to the first of these, Wright moves from the extension-determining claim to the conclusion that avowals are not expressive of any cognitive achievement. Moran finds this inference surprising. First, he notes that Wright introduces the idea of the judgement-dependent character of intentional state concepts by drawing an analogy with secondary qualities such as colour. He then points out that, supposing the concept of colour were judgement-dependent in this way, it "would certainly not follow that ... particular judgments of the color of something were not expressive of a cognitive (indeed perceptual) achievement of some sort, and were instead a matter of some kind of social concession" (Moran, 25). The same, he asserts, goes for our concept of intention. "So," he continues, "even if the relevant biconditionals for intention could be specified non-trivially and their *a priori* status secured, this would not serve to show that first-person

authority was not based on some kind of genuine cognitive advantage" (ibid.). In fact, according to Moran it is even worse than that for Wright. For, as he sees it, the case for extension-determination cannot "even serve the purpose of ruling out a perceptual model of introspection, as the color analogy shows" (ibid.).

This last claim is somewhat curious. Assuming that the *a priori* status of the concept of intention is secured by the extension-determining nature of our judgement about our intentions, then Moran must be arguing that an extension-determining second-order judgement can nonetheless involve "some sort" of cognitive achievement.[24] However, if he is claiming this, then the last sentence quoted above seems off-base. With regard to perceptual models of introspection, the traditional idea – which seems to be what he has in mind – is that the subject perceives a mental state that exists independently of any second-order judgement about it (it is both ontologically and conceptually distinct). Such judgements are extension-reflecting. So, contrary to what Moran claims, Wright's account does rule out such a perceptual model of introspection.

How about the claim that the judgement-dependent character of intentional state concepts doesn't explain the asymmetries between first- and third-person ascriptions? Take first the claim of privileged authority. According to Moran,

> Nothing in the analysis explains why there should be any difference at all in the application conditions of psychological concepts in first-person and third-person contexts. ... For we could specify a similar set of biconditionals as governing the application of psychological concepts to others. That is, we could specify C-

[24] In fact, this is what he will argue later on, when he discusses the relation between the first-person perspective, avowals and what he terms the "deliberative stance".

> conditions, competent ascribers, conceptual capacities, and so forth, in such a way as to make it an *a priori* matter that such ascriptions have a strong prima facie claim to truth. (Ibid.)

As an example he offers Davidson's theory of radical interpretation (and the principle of charity it includes), where it is argued that the possibility of interpretation requires that the interpreter have a solid grasp of the beliefs and other mental states of her interlocutor.

However, even if such a theory makes it an *a priori* matter that other-ascriptions must have a strong prima facie claim to truth, it is not of the same level or kind as that entailed by Wright's constitutive thesis. For one thing, according to Davidson there is a crucial difference between self- and other ascriptions, namely that the latter always involve interpretation (in the "radical" sense), while the former do not (Davidson: 2001c, 12-13; 2001b, 37; 2001e, 66). This allows for a possible source for error in the former that is absent in the latter. Similarly, one might argue that, according to Davidson, it is an *a priori* matter that our judgements about the world have a strong prima facie claim to truth.[25] The possibility of interpretation/mutual understanding requires it. But again, this is not the kind nor degree of authority we are thought to enjoy with respect to our avowals. And, Wright might argue, it is because third-person ascriptions are extension-reflecting, as opposed to extension determining, that this difference exists.

Turning to the question of immediacy, Moran writes:

> For all the biconditionals tell us, it could be that first-person ascriptions were only made on the basis of examining the evidence provided by one's own behaviour, but

[25] See 4.2e for a discussion of this claim.

> our convention dictated that we always privilege the person's own reading of that evidence as the best possible one. (Moran, 26)

Two points can be made here. First, one wants to ask: behavioural evidence of what? Presumably, of a first-order state. This suggests that Moran is contesting the idea that the *a priori* status of the biconditional requires the extension-determining conclusion after all (and not merely that cognitive achievement is consistent with extension-determination). For wouldn't the reading of evidence, which is presumably independent of our judgement about it, suggest that the judgement was extension-reflecting? Second, Moran says that, for all the biconditional tells us, first-person authority could be a matter of convention that has us privileging a person's own reading of the evidence. Is he suggesting that it could be shown to be *a priori* reasonable to suppose that avowals are extension-reflecting (based on evidence), and that the subject's own reading of that would be best? If so, on what grounds does he think this? As Wright points out, this is sufficiently counter-intuitive that some explanation is required.

Moving on to the final objection, Moran writes:

> [A]ny adequate analysis of the first-person would have eventually to get beyond the picture of "privilege" and concessions and say something about how the presumption of first-person authority expresses an ordinary rational *demand*, quite as much as it reflects any deference to the person's best opinion about his own state of mind. ... ("Do you intend to pay the money back?" "As far as I can tell, yes.") (Ibid.)

The point here is that Wright's appeal to "social concessions" fails to account for the sort of commitment unique to first-person authority. We don't merely defer to a subject's "best opinion," as Wright puts it (Wright: 1998b, 204); rather, we expect that in avowing, for example, an intention, she express a more robust commitment to the state of affairs reported.

It seems that with this objection Moran focuses on what might be considered a somewhat unfortunate use of terminology, one that perhaps leads him to misunderstand the basic thrust of Wright's view. Yes, according to Wright it is part of our treatment (and concept) of a rational subject that we "unofficially" grant to her the ability to make accurate judgements about her own mental states. But this is because this is what it is to be in such a state – mental states are partly constituted by those judgements. If so, then the response in the money example, insofar as it implies an extension-reflecting or "detective" understanding of avowals, would seem to be off the mark. That said, Moran might have a point here; given that all avowals are defeasible – contingent upon the (C) conditions being met – the response could be read as an acknowledgement of that fact. In other words, in spite of the positive-presumptive status of those conditions, one might still acknowledge the possibility that evidence that one has not yet recognised exists that would place the truth of the avowal in doubt.

However instructive it might be, a full discussion of Wright, and Moran's analysis of his view, would take us too far afield. For the purposes of this discussion, it is enough to note that Moran's critique introduces a couple of related ideas that will figure prominently in his own account of self-knowledge. As he sees it, Wright's account

manifests a general tendency to which philosophers often fall prey when discussing self-knowledge. Most discussions of self-knowledge exhibit what he calls an overly "theoretical" orientation; it is assumed that if there is to be self-knowledge, it must resemble the situation of making judgements about the external world (Moran, 27). As mentioned above, this limits the understanding of introspection to an inner-perceptual or spectatorial model, and self-knowledge to the formation of belief about a static realm of mental facts. And, when that is found wanting, the deflationary conclusion is thought to follow. This comes at the expense of the consideration of what significance the kind of self-knowledge expressed by avowals has for the subject, the understanding of which, Moran argues, is key to arriving at a satisfactory explanation of self-knowledge and the asymmetries that distinguish it from third-person judgement. As he puts it, "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 awareness" (ibid., 28).

3.1.d Self-Constitution and the Supposed Insubstantiality of Self-Knowledge – Moran on Taylor

Showing how self-consciousness has specific consequences for the object of awareness will be a key component of Moran's own account of the substantiality of self-knowledge. However, he notes that some philosophers have taken this idea to be incompatible with what Moran calls an ordinary realism about mental states and substantial epistemology of self-knowledge. For example, he refers to Charles Taylor (1985), who argues that the

"traditional theory of consciousness as representation" does not reflect the dynamic relation between a subject's first-order mental states and her judgements about them. According to the ordinary notion, "representations are of independent objects. I frame a representation of something which is there independently of my depicting it, and which stands as a standard for this depiction" (Taylor, 100). However, this picture doesn't match the situation with respect to self-interpretation, where reflection on one's mental state may inform its character:

> Formulating how we feel, or coming to adopt a new formulation, can frequently change how we feel. We could say that for these emotions, our understanding of them or the interpretations we accept are constitutive of the emotion. ... And that is why the latter cannot be considered a fully independent object, and the traditional theory of consciousness as representation does not apply here. (Ibid.)

If the way in which a subject conceives of her own mental state may be at least partly constitutive of what that state is, then it would seem as though there is no room to speak of that state having a fully objective existence, that is, one fully independent of the subject's own judgement of it.

Moran also rejects what Taylor calls the "traditional theory of consciousness as representation" (what Moran terms the spectatorial model of introspection). But he disagrees with what he takes to be Taylor's deflationary conclusion that the potentially constitutive/transformative power of (supposedly) higher-order judgement Taylor describes undermines ordinary or commonsense realism about mental states and self-knowledge. Referring to an example of Taylor's (ibid.), he agrees that when, upon

reflection, one comes to a new understanding of one's situation that holds a feeling of guilt to be unwarranted, this may very well lead to a change in that attitude (but it may not – more on this below). This, he says, is to be expected, assuming one is a rational agent. Any attitude partly grounded on judgements about the world ought to be subject to rational criticism. If one's understanding of the facts upon which one's feeling of guilt is based changes, leading to a second-order belief that a change in the first-order state is warranted, then that state should be sensitive to this new understanding and should change accordingly. However, it does not follow that there is no independent fact of the matter about the subject's mental state prior to the reflection and formation of the second-order belief that some sort of change was in order.[26] Consequently, there is no reason why this should undermine either commonsense realism or "consciousness as representation" regarding mental states (Moran, 53-54).

3.1.e Self-Knowledge is a Rational Requirement

According to Moran, Taylor's deflationary conclusion is another instance of the tendency to equate the possibility of a substantial account with the applicability of the observational model (ibid., 37-38). But in fact, as Moran sees it, the dynamic relation between first-order states and second-order beliefs about them that supposedly leads to the deflationary conclusion is what, when understood in terms of its role in rational

[26] In fact, it is presumed that one already has a correct second-order belief about the content of the first-order state that is the object of second-order reflection.

agency, reveals the substantial character of properly first-personal self-knowledge (that is, the kind of self-knowledge that is unique to the first-person perspective).

Moran distinguishes between what he calls *theoretical* and *deliberative* self-knowledge (ibid., 55-59). The former is essentially third-personal, in the sense that it is restricted in scope or perspective to the description of the psychological facts about oneself. Theoretical inquiry into one's states ends with a second-order belief about the content and/or quality of them. Such inquiry is in the mould of traditional observational theories of self-knowledge (or what Taylor calls consciousness as representation), in that one's relation to the object known is like that in ordinary empirical knowledge. While it may be that such knowledge typically derives from observation of one's behaviour or thoughts, the understanding (explanation) of which requires interpretation (as, for example, in the therapeutic situation), this is not its key feature. We could, at least for the sake of argument, suppose that one had some sort of mind-reading faculty that provided for immediate and reliable access to one's mental states. Or, one might be constantly assailed by true thoughts regarding one's mental states. In other words, such knowledge could, in principle, be immediate and reliable and still not count as first-personal at all. For what defines this sort of self-knowledge is the total independence of the object known (the subject's first-order states) from the knower (the self-conscious subject). In theoretical self-knowledge, the subject's relation to her mental states is one of alienation; as a mere observer of the psychological facts, the state known and knowing state are, as Moran puts it, cognitively isolated from each other, with the latter having no impact on the former (ibid., 60).

Contrasted with this is deliberative inquiry, which is essential to the first-person perspective. In this form of inquiry, one adopts what Moran calls a deliberative stance toward oneself (ibid., 59). Judgements made from this perspective are akin to practical reflection, in that they end not merely in belief about the content or character of a first-order mental state, but in a commitment to, or endorsement of, the content of that state. Such inquiry conforms to what Moran terms the *Transparency Condition* (ibid., 67). That is, questions such as 'Do I believe (desire, intend, etc.) that *p*' are "transparent" to (but not equivalent or reducible to) questions regarding the subject matter of *p* itself. The idea is familiar from Gareth Evans, who describes it thus:

> in making a self-ascription of belief, one's eyes are, so to speak, or occasionally literally, directed outwards – 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 in answering the question 'Will there be a third world war?' I get myself in a position to answer the question whether I believe that *p* by putting into operation whatever procedure I have for answering the question whether *p*. (Evans, 225)

In adopting the deliberative stance – in conforming to the Transparency Condition in arriving at her self-ascriptions – the subject is guided by a commitment to 'rational authority', or the authority of justifying, as opposed to explanatory, reasons in determining her beliefs, desires and other first-order mental states. Commitment to the Transparency Condition lies at the core of Moran's understanding of the link between self-knowledge and rational agency, as the subject exercises rational control over his

mental life to the extent that he undertakes this commitment. It is also key to the explanation of the distinctive features of authoritative self-knowledge. To quote Moran,

> The authority the agent does speak from, as well as the fact that his declaration is made without observation of himself both stem from the fact that the person's own relation to his attitudes and his intentional actions must express the priority of justifying reasons over purely explanatory ones. In this, his position is fundamentally different from that of another person, such as a rationalizing interpreter, who is seeking to explain either his actions or attitudes. The reasons that explain an action are states of mind of the agent, which may be veridical or mistaken...But naturally this is not the agent's own relation to his reasons, which must be guiding or justifying reasons, and which are facts distinct from and independent of his beliefs. (Moran, 128)

The commitment to justifying reasons essential to the first-person perspective also accounts for the non-evidential character of avowals. Since the speaker commits to the primacy of justifying reasons, there is nothing more for her to do once a conclusion is reached (such as introspect to check that she really does believe) to make that belief her own. As Moran puts it, "the goal of deliberation, whether practical or theoretical, is conviction, about what to do or what to think" (ibid., 131). Consequently, a non-evidential access to one's beliefs is a basic requirement of rational agency; a failure of transparency in one's deliberation equals a failure of self-knowledge, a failure to reach a fully conscious or first-personal state of knowledge of one's mental life.

Seen from this perspective, the question of what mental state I have is answered by my judgement about the facts related to whatever is the object of the mental state in question. This is a uniquely first-person phenomenon – only my interpretation of my mental state will be self-constituting. Another's interpretation (for example, a psychoanalyst's) may influence my understanding, but ultimately this will depend upon my being fully convinced (for the right reasons) of the truth of the interpretation of that view. So, to take the earlier example, if my reflection on the object of my resentment (over, say, a comment made by a well-meaning friend) leads to the conclusion that that reaction is unwarranted, then that should lead to a new awareness and transformation of that state in line with my commitment to, or endorsement of, the ("external") facts as I now judge them to be. The subject who is incapable of this, who is only able to take an empirical or third-person perspective on herself, fails to exercise the requisite authority with respect to her mental states and is not fully 'conscious'. Her inability to move, where appropriate, from a third- to first-person point of view, from self-observation to "genuine" avowal, in the expression of her second-order belief constitutes a failure of reflective rationality and self-knowledge.

In light of these considerations Moran argues that we need an expanded understanding of first-person authority, one that goes beyond the asymmetries of immediacy and reliability. It is when we consider the notion of authority in terms of the subject's exercising control over her mental life by submitting to the authority of what she judges are the best reasons to believe (desire, intend, etc.) that *p*, that the immediacy and reliability characteristic of the kind of self-knowledge philosophers have traditionally

been interested in are explained. In approaching the question of what she believes (desires, intends, etc.) from a deliberative perspective, a subject is committed to whatever conclusion she eventually draws. But, Moran notes, this is not to say that all "genuine" *avowals* – that is, those self-ascriptive judgements that obey the Transparency Condition – arise out of such an explicit process of deliberation. Rather, it is only necessary that in sincerely uttering, "I believe that *p*," the subject endorse the embedded state (the belief that *p*), where that endorsement reflects a commitment to the Transparency Condition such that the subject would alter that belief if deemed appropriate in light of her subsequent judgement of the facts upon which it is based.[27] These are the essential features of the deliberative stance.

We find, then, that the explananda that have usually been of interest in discussions of self-knowledge (namely the immediacy and reliability thought to characterise avowals) are neither the only, nor essential, distinguishing features of it. The immediacy and reliability of avowals are consequences of (but not necessarily restricted to) the subject's ordinary exercise of her capacity to "obey" the transparency condition when making judgements about her own mental states. Put another way, they are a natural consequence of self-ascriptions made from the deliberative or first-person perspective (the capacity to avow being the mark of the first-person).

[27] As Moran puts it, an avowal is "the expression of one's own present commitment to the truth of the proposition in question" (Moran, 86).

3.1.f The Transparency Condition and the Double Expressive Character of "Genuine" Avowal

Moran argues that the key to explaining the asymmetries associated with self-knowledge, as well as its substantial character, is in understanding in what first-person authority essentially consists. But as he remarks, this defining feature of avowals – the Transparency Condition – is just what some philosophers have pointed to in support of a deflationary account. If an avowal of, say, 'I believe it will rain' is transparent to 'It will rain' – that is, if they are based on the same considerations – then it follows that the former is really just another way of expressing the latter, albeit in a more or less guarded fashion. In other words, a superficial grammatical difference masks what are essentially the same judgements. As Moran puts it, "in the first-person present-tense use, the verb phrase 'I believe' does not in fact have any psychological reference, but is instead a mode of presenting the relevant proposition" (ibid., 71). So, what we really have are two different ways of stating the same judgement that *p*. And, if this is so, then what are (mistakenly) taken to be self-knowledge claims – that is, expressions of second-order beliefs about first-order mental states – are simply claims about, for example, the weather.

According to Moran, this is very similar to an expressivist account of self-ascriptions. As he understands it, expressivism is a claim about self-ascriptions in general, namely that they are not assertions and thus do not serve to describe, or report on one's mental state, but rather only express it. On this view, an utterance of "I have a headache" is essentially the same as a yelp or the exclamation "ouch"; that is, it is a behavioural

expression of pain, and has only an apparent reference to the subject. And, lacking such a self-referential assertoric status, it follows that self-ascriptive utterances cannot be self-knowledge claims. For, as Moran points out, "[o]nly a report of one's state, or some other fact-stating utterance, can be something that is true or false, justified or unjustified" (ibid., 102). If self-ascriptions do not report, but merely express, one's mental state, then no epistemological questions regarding the grounds for, or reliability of, the claim may arise.

The deflationary conclusion is thus grounded on the negative claim that self-ascriptions serve to express, but not report, one's mental states. What support is there for such a claim? Moran argues that, generally speaking, an utterance may serve multiple functions. For example, "The brakes don't work" may both state or report a fact (that the brakes don't work) and express the speaker's belief (and perhaps panic and/or surprise, depending on the context). But, Moran asks, why should we think that, when it comes to self-ascriptions, these two categories are mutually exclusive? In fact, he maintains, there are good reasons for thinking the opposite:

(1) Expression is a general category, "encompassing both verbal and non-verbal behaviour, as well as both the overt, deliberate declaration of one's state and the involuntary manifestation of it" (ibid., 104).

(2) "Understood this way, reporting or describing one's state is a particular *way* of expressing or manifesting it. It is a special way, involving a judgement about one's state of mind and the special responsibilities involved in *asserting* that judgement" (ibid.).

(3) "The denial of this general possibility ... is therefore more properly put forward as the claim that first-person expression of pain or belief are *mere* expressions; that is, not to be included among the verbal expressions that are also assertions or reports of one's state" (ibid.).

(4) However, "[t]his general denial loses all plausibility when we recall that the category of reports is quite broad, broad enough to include ascriptions made on any basis whatsoever, as well as those made on no basis" (ibid.).

(5) Thus, what Moran calls mere attributions (those self-ascriptions the subject fails to endorse), made either on the basis of "third-personal" self-interpretation (that is, on the basis of the subject's own interpretation of her thoughts and/or behaviour) and/or the authority of another (for example, a therapist), must count as reports and not mere expressions.

If this is so, then expressivism, as a general thesis about self-ascriptions, is false. But, Moran asks, what about an expressivist reading of those self-ascriptions – that is, "genuine" avowals – that obey the Transparency Condition? Might at least they not be "disguised" expressions of first-order mental states, with only the appearance of second-order assertions about one's mental life? Not as he sees it. He writes:

(6) What conforming to transparency comes to is the commitment that beliefs I call my own are beliefs I can endorse as true. But that commitment is internal to the very concept of belief and cannot itself annul the prima facie reference to oneself in a statement like "I believe it has stopped raining." Any understanding of belief that provides for the minimal idea that believing involves "holding true" will

> entail that it is at least possible to announce one's belief by reporting on the truth as one sees it. If my intention is to report on my belief as such, and I know … that my belief about X is what I hold true of X, then my intention will not be thwarted if I make this report by considering what is true of X. (Ibid., 105)

My intention to report my belief about X – that is, say something "with the intention of telling another person my thoughts, beliefs, and feelings" (ibid., 71) – may be realised by my reporting on the truth as I see it; that is, by expressing my first-order belief about X, which may be expressed in a self-ascription. But, Moran maintains, this orientation does not annul the prima facie reference to oneself, which is to say it does not cancel its status as an expression of second-order belief about what one holds to be true of X. In other words, avowals should be understood as having a double expressive role, expressing both the speaker's first-order state and second-order belief about it.

In defending this view Moran points out that to deny it in favour of the expressivist account would imply what he sees as the highly implausible denial of the possibility of non-attributional self-knowledge. So when during therapy an analysand moved from the mere attribution of a mental state to its avowal, that would not count as a development of her self-knowledge. For, on this view, we would be left with the "perverse idea that only an essentially third-person perspective on oneself could count as a vehicle for self-knowledge" (ibid., 106).

Perverse or not, this, or something close to it, does seem to be the consequence of the expressivist view. But describing it thus does not count as an argument against it. What, in that respect, has Moran offered to persuade us against it? Recall the claims (1) to (5)

above, from which it is concluded that expressivism, as a general thesis about self-ascriptions, is false. As Moran understands it, the expressivist argues that a self-ascription never counts both as the expression of a (first-order) mental state and second-order report of it. But, he points out, an utterance such as "The brakes don't work" may both report a fact (that the brakes don't work) and express the speaker's (first-order) belief that the brakes don't work. Furthermore, as he notes in claim (5) above, at the very least there are self-ascriptions that, based as they are on the interpretation of behavioural evidence, clearly are assertions about one's mental state. So, it is concluded, it cannot be that self-ascriptions *never* serve as reports about one's psychological state.

This may be so; however, it does nothing to boost Moran's positive claim about the double expressive character of avowals. For the defender of the expressivist position is still free to argue that those ascriptions that conform to the Transparency Condition are reports even while serving an exclusively first-order expressive role.

This is where the argument presented in passage (6) comes in. Moran defends the double expressive character of avowals by appealing to the speaker's intention to inform her interlocutor of her belief about X, which, in accordance with the Transparency Condition, she accomplishes by considering the facts about X. But, if "it is at least possible to announce one's belief by reporting on the truth as one sees it," then one could fulfill one's intention to report one's belief about X simply by expressing it. In other words, the extra level of assertion – that one asserts that one has the state one expresses (here, a belief about X) – that is said to belong to the avowal is unnecessary to fulfill one's intention. So if the issue were to be decided on the grounds of this argument alone,

the proponent of the expressivist view might invoke the power of Occam's razor to excise the superfluous level of belief.

It does seem that further argument is required to settle this issue, and I will return to the examination of this matter in the next chapter.[28] But say for now we accept Moran's conclusion. Several other potential difficulties for his view remain. Part of his aim is to explain the relation between the deliberative perspective and the first-person, or how the deliberative stance and the kind of self-knowledge that arises from it are essential to the first-person perspective and rational agency. But an examination of what this involves puts into question the substantiality of the self-knowledge it is invoked to support – there is some question as to whether or not the kind of self-knowledge that he argues is intrinsic to the first-person is substantial at all. Moran argues that we can see how self-knowledge intrinsic to the first-person is substantial by understanding the connection between that kind of self-knowledge and rational agency (how it is a rational requirement). But what is required is that one's self-ascriptions obey the Transparency Condition, that is, derive from one's first-order judgement of the reasons for the state reported. If so, then it would seem that the substantial cognitive achievement involved in an avowal of 'I believe that *p*' would be the same as that involved in a sincere assertion of *p* itself. But then it would seem that avowals would lack any substantial cognitive achievement of their own.

[28] See 4.1.e, as well as the discussion of Bar-On's own appeal to 'dual-expressivism' in 4.1.h.

3.1.g "Endorsement" and the Immediacy and Reliability of Avowals

Let's look again at the avowal of belief. To reiterate, the idea behind the deliberative stance is this: I realise my intention to report about what I believe about *p* by considering the facts about *p*. Thus, when I am finished with this deliberative process, there is nothing I need do to make that belief my own when I report it. This accounts for the immediacy and reliability of such "explicitly deliberative" avowals. However, as Moran notes, we do not always arrive at our beliefs, desires, and other first-order states through such an explicit deliberative process. In such cases, it is sufficient that one's self-ascription be made in a "deliberative spirit" – that one endorse or be committed to the state one self-ascribes as answerable to one's best reasoning about its subject-matter. In this sense, the first-personal self-knowledge expressed in avowal is an expression of my rational agency or authority; and it is this commitment that defines such self-knowledge and distinguishes it from the third-personal or merely attributional variety (in other words, that made by the first-person from the theoretical perspective). To summarise, an avowal is a self-ascription that is made from this deliberative perspective. In avowing her mental state to another, the speaker seeks to report that state by expressing it, where the state she expresses is a state she commits to on the grounds (either implicitly or explicitly) of what she takes to be its reasonableness (that is, the reasons she sees for it).[29]

In explicating the commitment that distinguishes "genuine" first-person self-ascriptions (avowals in Moran's sense of the term) from "mere" third-personal self-

[29] In other words, it expresses her commitment to her state being determined by her reflection on the object of the attitude, even if she has not yet engaged in such deliberation (Moran, p. 85).

attributions Moran presents what might be described as a more refined articulation of first-person authority. Typically, discussions of self-knowledge and first-person authority have focused on the explanation of the immediacy and reliability of a certain class of self-ascriptions (those not based on observation or inference), and how those self-ascriptions may or may not count as knowledge. Moran does address these issues, but for him the primary matter of interest is how we should understand the relation between self-knowledge and the first-person; he seeks to show that what kind of relation a subject bears to, or form of awareness she has of, her own mental states is definitive of the subject as rational agent. And with regard to the explanation of the immediacy and reliability of self-ascriptions, it turns out that these are not the defining features of first-person authority as is traditionally thought. In his discussion of explicit deliberation, Moran points out that in such cases the subject's attention is oriented "outward". He comes to awareness of his mental state – for example, his belief about p – through his consideration of the facts about p, which he then reports by expressing the first-order belief he has formulated on the basis of that deliberation. It is the direct expression of the first-order belief embedded in the report that explains the immediacy and reliability of the avowal (which, again, counts as an avowal in virtue of the second-order judgement conforming to the Transparency Condition). This is why avowals that are not the result of explicit deliberation share these features – even without explicit deliberation, the subject expresses the state she reports, the expression of which, arising as it does from the deliberative perspective, is also an expression of her rational agency.

However, contrary to what Moran supposes, nothing here rules out the possibility of immediate and reliable self-ascriptions of states to which I am not committed in the sense avowal requires. That is, a self-ascription may be a mere attribution (in other words, one that I do not endorse), yet immediately and reliably made. Take the avowal of desire. Imagine an out of control gambler who desires to quit the tables. He arrives at this desire through the consideration of how gambling is ruining his life. His desire to quit is formed out of his deliberation on the facts of his situation. But now suppose that he continues to correctly self-ascribe a desire to gamble, in spite of his understanding of all the reasons against having that desire. As Moran might say, in this case, "there is still work to do," in the sense that he has a continuing competing desire to gamble that remains impervious to his deliberations on the desirability of gambling. This self-ascription is an expression of third-personal self-knowledge in the sense that it is as a "mere" attribution, something that he does not endorse. Still, insofar it expresses both his first-order desire to gamble (however dissociated from his better judgement), as well as his second-order belief that he has such a desire, it remains an immediate and true self-ascription of his current mental state.

A further difficulty for Moran's account arises from considerations of self-ascriptions that are not plausibly thought of as conforming to transparency. For example, my self-ascription of "I have a headache" is not an ascription of something that we think of as deriving from any sort of deliberation. Yet, self-ascriptions of such phenomenal states do seem to be equally secure and immediate as "genuine" first-person ascriptions (that is,

avowals). It would seem, then, that Moran should offer some additional account of how we are able to so reliably and immediately produce such second-order beliefs.

3.1.h A Substantial Account of Self-Knowledge?

Given this explanation of the immediacy and reliability of non-observational non-inferential self-ascriptions, we may ask to what extent Moran's explication of "genuine" first-person authority counts as a substantial account of self-knowledge. As mentioned above, Moran argues that one feature a substantial first-person self-knowledge claim ought to have is that "its truth conditions be in some way independent of the making of the judgement" (ibid., 18). This, he maintains, "is a form of cognitivity that any account of introspection as a source of knowledge would seek to preserve" (ibid.). This observation is made in response to Tyler Burge's account of basic self-knowledge, where simply making a judgement (e.g., "I judge, herewith, that writing requires concentration") guarantees its truth. The truth of the self-ascription is guaranteed because the state reported is also expressed. However, this is just what Moran argues for avowals, where one realises one's intention to report one's belief, desire or other first-order mental state by expressing it.

According to Moran, he is "pursuing an understanding of self-knowledge that would make sense of both success and failure in introspection...and thus accommodate some independence of awareness and the objects of awareness" (ibid., 20). In what sense is this provided for in his account of avowal? Given that avowals are defined as self-ascriptions

that obey the Transparency Condition, then it would seem that they are necessarily true. They are necessarily true in virtue of the commitment they express to the truth of the state reported. However, while false avowal may be a contradiction in terms, one may have a state that one fails to avow. Many of the examples Moran offers to illustrate instances of self-knowledge involve the therapeutic situation, where a subject comes to a true belief about her psychological state through analysis of her behaviour and thoughts. For example, on such a basis an analysand may come to the true conclusion that she feels resentment toward a family member; however, when she examines the facts regarding how she has been treated, she can see no reason for it. The claim is that this is not genuine self-knowledge, because while she may correctly attribute a state to herself on the grounds of her self-observation, she is not able to self-ascribe it on the basis of her understanding of the facts regarding the parent's treatment of her.[30] Two possibilities present themselves here. One is that she has the belief, but that it is unjustified and warrants change. The other is that she must re-examine the facts to see if, after all, they do justify her belief and feeling. Either way, the failure to avow indicates that something is amiss in her relation to her first-order state. Conversely, according to Moran, her coming to avow her state constitutes a development in her self-knowledge, a coming to first-person awareness of a state that was previously inaccessible in this regard.

Arriving at this first-person awareness (as he puts it) seems to be the substantial cognitive achievement unique to genuine self-knowledge that he has in mind. However, again we may ask: in what sense is coming to an understanding of the facts regarding a

[30] As Moran puts it, "she will affirm the psychological judgement 'I believe that P', but will not avow the embedded proposition P itself" (Moran, 85).

certain subject matter that leads to the formation of a first-order attitude, which one then reports on through its expression, an instance of self-knowledge? As Moran points out, the focus of my judgement and commitment in the avowal 'I believe (intend, desire, etc.) that *p*' is the same as in my utterance of '*p*'. So it is difficult to see how there is a difference in the type of cognitive achievement involved in the two judgements here. The only difference between the two is the additional second-order belief said to be expressed in the avowal. But given that there is no possibility of discord between these two levels, the addition of this assertoric self-referring role hardly seems to supply the kind of cognitive achievement that would make for a substantial epistemology of self-knowledge.

I have suggested that, at the very least, Moran has failed to make the case for a *substantial* epistemology of self-knowledge. However, might avowals remain instances of insubstantial self-knowledge? The case for Moran's view of the knowledgeable status of avowals rests on the strength of his argument for their double expressive character. In the following chapter we will take a closer look at the expressivist position, in the process of which the plausibility of the double expressive claim will be assessed. But before that, I turn to the examination of Davidson's non-epistemic explanation of self-knowledge. By considering what turns out to be an ineffective expressivist challenge to Davidson's account of self-knowledge, we will see the need for a more nuanced formulation of the expressivist view, the explanation of which will be undertaken in Chapter Four.

3.2 Davidson: Self-knowledge and Semantic Authority

I have already looked briefly at Davidson's account of first-person authority in the earlier discussion of Tyler Burge's view of self-knowledge. In the second part of this chapter I offer a more detailed examination through a consideration of some criticism offered by P.M.S. Hacker. After examining Hacker's criticisms, and arguing that they betray a fundamental misunderstanding of Davidson's position and thus miss the mark, I will then point out what might be described as a non-fatal flaw in Davidson's account, namely that, on its own, Davidson's explanation of semantic authority cannot explain the asymmetrical character of our self-ascriptions. This will set up the discussion of expressivism to come in Chapter 4.

3.2.a Subjectivism and the Denial of First-Person Authority

As touched on above (Section 2.2.f), Davidson argues that a proper account of semantic authority (that is, our authority with respect to the meanings of our own words) provides us with an explanation of our capacity to reliably and immediately self-ascribe our mental states. To further explain Davidson's view we can turn to his critique of Hilary Putnam's version of semantic externalism, which Davidson takes to undermine first-person authority. Putnam argues that the meaning of the words are determined in part by the nature of the objects to which they refer (thus his claim that meanings are not completely "in the head" (Putnam, 227). We have already seen that Davidson partly shares this externalist view of meaning. How does he understand Putnam's version to compromise knowledge of one's own mental states? Putnam distinguishes between two sorts of

psychological states: (1) 'narrow' states, that is those described solely in terms of the isolated subject (including her physical make-up), and (2) 'wide' states, namely those that include the subject's relations to the object that the subject's propositional attitude is about (ibid., 220). Putnam argues that two persons could be identical in every way with respect to (1) and yet unknowingly mean different things by the words they use (ibid.,227).

Putnam illustrates this situation with his Twin Earth example, where we are asked to imagine two physically identical persons living on separate worlds which are also identical in every way except that on Earth the stuff called 'water' is composed of H_2O, while on Twearth it is composed of XYZ (with all other phenomenal characteristics – for example, the taste and texture – being the same) (ibid., 223 ff). Both speakers, ignorant of the chemical composition of the respective substances, use the same (sounding) word 'water' to refer to the respective liquids. The problem for first-person authority as Davidson sees it is that

> if people can (usually) express their thoughts correctly in words, then their thoughts – their beliefs, desires, intentions, hopes, expectations – also must in part be identified by events and objects outside the person. If meanings ain't in the head, then neither, it would seem, are beliefs and desires and the rest. (Davidson: 2001b, 18)

When each speaker in her home world sincerely says, "That's water" in reference to the liquid, she expresses a true belief. In the case of the Earthling, she expresses the belief that what stands before her is a glass of water (H_20). The Twearthling expresses the belief

that before her is a glass of twater (XYZ). The two speakers are identical with respect to their narrow states; however, because they unknowingly differ with respect to their wide states, they mean different things by the use of their words. But because this difference in their wide states is undetectable by them, it follows that they do not have a full grasp of what they mean and believe. If the meanings of the words that a speaker uses to express her beliefs are not completely in the head (exhausted by one's narrow state) – that is, are determined in part by the objects to which she is causally related – and the subject is ignorant of some aspect of that external object, then she must have only partial understanding of what she believes. For we can only fully know the meaning of our words and what we are claiming with them if narrow and wide states correspond. However, this is seldom, if ever, the case; therefore, because we don't have a complete grasp of the meaning of our words, we lack first-person authority with respect to our beliefs and other mental states.

According to Davidson this view of meaning that he attributes to Putnam presupposes a mistaken view of the subjective whose influence persists even when its problematic nature is revealed. It is the idea that likens the mind to an inner theatre, where the conscious self observes or grasps mental objects that appear before it. From early on this understanding of mental life led to sceptical problems concerning the veracity of our knowledge of the world (including other minds), namely how one could know that the objects that exist in or before the mind (which we *can* know) accurately represent that which is 'out there' in the external world. This picture has continued to inform talk of what it is to be in a state of mind with respect to propositional attitudes (Davidson:

2001b, 34-35). Davidson argues that this view, combined with a mistaken conception of what constitutes word meaning and the content of belief, is what prevents the resolution of the problem of first-person authority.

3.2.b Davidson's Critique of Putnam's Semantic Scepticism

Davidson agrees with Putnam with respect to the role our interaction with objects in the world plays in the formation of mental content. It is the divide between the narrow and wide states – which, he argues, is a residue of the Cartesian view of mind – that he says we should do without. According to Davidson, Putnam is committed to the sceptical conclusion regarding first-person authority because of two assumptions:

(1) Understanding the meaning of a word, and thus being able to use it correctly, involves having something in or before the mind that the mind is able to grasp. It is in having this mental object before the mind and grasping it that the subject is guided in her correct use of that word. So on this view, it is the subject's understanding of a pre-existent meaning that explains how she correctly uses a word.

(2) Since the external object to which the mind is causally related determines the meaning of the word, then whatever determines the nature of that object must be grasped by the subject's mind if she is to know what state of mind is expressed through the use of that word (ibid., 35).

In effect, Putnam has gone half way toward a non-Platonic conception of meaning. On the Platonic view, we understand the meaning of a word and the concept it signifies (for

example, 'justice') when we grasp the Form of Justice that defines it and which determines the way the word ought to be used. In grasping the Form we acquire the knowledge that allows us to use that word as it ought to be employed as dictated by that independent norm. While Putnam has exited the ideal world of Forms – bringing the norms that determine meaning down to earth, so to speak – he keeps to the idea that understanding the meaning of a word involves grasping a norm that "like a Platonic meaning that is just waiting there for the learner to grasp" and which, upon acquisition, guides the speaker in her use of it (Davidson: 2003, 692).

As Davidson sees it, eschewing this norm- or rule-governed picture of linguistic understanding and the mental opens a path to understanding the nature of meaning that reconciles first-person authority with the view that the identity of a belief is determined in part by the object the belief is about. More specifically, instead of taking meaning to determine use, we ought to see our use of words as determining their meanings. Davidson fleshes this idea out as follows. First, there is the semantic externalist claim that the meaning of a person's words "depends in the most basic cases on the kinds of objects and events that have caused the person to hold the words to be applicable; similarly for what the person's thoughts are about" (Davidson: 2001b, 37). Added to this is the regularity thesis, namely the claim that whatever objects and events a person regularly applies her words to – in other words, whatever way they are regularly used – gives them the meaning they have (and her thoughts the content they have as expressed by her use of those words).

These two ideas ground his account of first-person authority. The explanation is completed with the observations that (i) as long as a speaker knows that she holds true the sentence she utters (that is, is sincere), and (ii) knows what her words mean (as determined by the way she consistently uses them), then she will know what she believes. With this, we can see how Davidson thinks the asymmetries are explained. Assuming sincerity, a speaker need not appeal to evidence, as others must, to know what she believes because the way in which she regularly uses her words constitutes what those words mean (and thus the content of her belief as expressed through the use of those words).

This is not to imply that speakers must be infallible with respect to their self-ascriptions. For example, self-deception remains a possibility under such an account. However, what is impossible, Davidson maintains, is that a speaker could be generally mistaken. "The reason," he says, "is apparent: unless there is a presumption that the speaker knows what she means, i.e., is getting her own language right, there would be nothing for an interpreter to interpret. To put the matter another way, nothing could count as someone misapplying her own words" (Davidson: 2001b, 38). In other words, it is a requirement of interpretability that speakers must be generally authoritative in their knowledge of what it is they mean and believe. But, it should be noted, for Davidson this authoritative knowledge is a form of linguistic know-how, and not a body of propositional knowledge about the meaning of words.[31]

[31] A clarification may be in order here: to say that semantic authority is grounded on a form of "knowing how" as opposed to "knowing that" is not meant to suggest that Davidson thinks that propositional knowledge is not involved in successful communication. As he writes in the

3.2.c P.M.S. Hacker's Expressivist Critique of Davidson

This approach to resolving the problem of first-person authority has come under criticism from various quarters. P.M.S. Hacker, for one, argues that Davidson's theory fundamentally misrepresents speech and how it is understood. While he supports Davidson's rejection of the Cartesian idea that beliefs are objects before the mind, as well as the claim that one's knowledge of one's beliefs is indubitable or incorrigible, he goes on to remark disapprovingly that Davidson remains within the "field of force" of the traditional subjectivist paradigm. This is because he "accepts as a datum that 'a person normally knows what he or she believes'" (Hacker: 1997, 287). Hacker rejects what he calls this 'cognitive assumption' (ibid.) on the grounds that first-person present-tense avowals of belief are not usually assertions or reports (in other words, knowledge claims) about one's mental state. On Hacker's view, which he regards as a form of Wittgensteinian expressivism, when I utter 'I believe that *p*', there is usually no epistemic use that attaches to the 'I believe that' part of the utterance (ibid. 291 ff). But if this is so, then there is no asymmetry in authoritive knowledge to be explained between first- and third-person ascriptions as Davidson frames it.

Appendix of *Truth, Language, and History*, "Knowing a language is, in some respects, like knowing how to ride a bicycle. In both cases ... we talk of knowing how, and in neither case is it necessary or common to know a theory that explains what we do. But there are also striking differences. There are endless things a speaker or interpreter must know: the truth conditions a hearer will probably take her utterances to have, the truth conditions that most of the sentences she hears will have, relations of entailment, contradiction and evidential support among sentences.... Bicycle riding requires no propositional knowledge at all" (Davidson: 2005, 325).

Hacker acknowledges that his claim may seem curious, for, as he points out, in sincerely saying 'I believe that *p*' I am not ignorant of my having that belief. "Does it not follow that when I believe, etc., that *p*, then, at least normally, I know that I do" (ibid., 291)? He maintains that it does not. Like Davidson, he points out that our claims to knowledge of our mental states are not based on any appeal to inference or observation. Hacker thus takes the immediacy of our self-ascriptions to be grounds for denying that they state or express knowledge of the ascribed mental state ibid., 290-291). However, unlike Davidson, he thinks it follows that one's present-tense self-ascriptions of one's mental states are about those states in appearance only; in other words, for Hacker there is no difference in assertoric content between a speaker's utterance of 'I believe that *p*' and '*p*'. As evidence he argues that the evidence one has for the self-ascription of *p* just is the evidence one has for the claim '*p*' (ibid., 290). So Hacker takes the transparency of avowals that was so central to Moran's account of the first-person perspective to be grounds for counting "I believe that *p*" as a semantically redundant formulation of "*p*".

What, according to Hacker, leads Davidson to make the mistaken cognitive assumption? In his critique Hacker emphasises Davidson's assertion that first-person authority is a necessary presumption for the possibility of communication, that it is a requirement of interpretability. He concludes that Davidson's argument for first-person authority takes the form of a transcendental deduction. He summarises it as follows:

> We know that we communicate with one another. It is a requirement of communication that there be a presumption that the speaker knows what he means by his utterances. But if he knows that he holds true the sentences he utters and

> knows what he means, then he knows what he believes. So there is a presumption, essential for the possibility of interpretation, and hence of communication, that a speaker knows what he believes when he avers that he believes something. (Ibid., 290)

On the surface this is an accurate synopsis of Davidson's position; each of these claims is in keeping with statements Davidson makes. That said, they tell only part of the story, since Davidson provides an independent argument explaining why what he calls our knowledge of the meaning of our words is more than a 'mere' presumption. As I shall argue below, it is Hacker's misunderstanding of this argument that leads him to misconstrue Davidson on first-person authority.

To see where he goes wrong, we need to turn to his rejection of the distinction Davidson draws between the ways in which a speaker and hearer know the beliefs expressed by an utterance. The basic difference regards the nature and role (or lack thereof) of interpretation in understanding. Davidson argues that the potential difference in word use between speaker and hearer means that, as he puts it,

> there can be no general guarantee that a hearer is correctly interpreting a speaker; however easily, unreflectively, and successfully a hearer understands a speaker, he is liable to serious error. In this special sense, he may always be regarded as interpreting a speaker. A speaker cannot, in the same way, interpret his own words. (Davidson: 2001c, 12)

Hacker rejects this, arguing that while some misunderstanding can be described as misinterpretation, it does not follow that correct understanding consists in correct

interpretation. In fact, it is just the opposite – in normal cases of successful communication, interpretation presupposes understanding. This is because for Hacker the concept of interpretation should be limited to that of paraphrase or translation that, unlike understanding, is a process or activity.[32] He writes:

> Typically, understanding the utterances of others involves no antecedent process of interpreting what was said in other words which are more perspicuous, since most utterances in context are already perspicuous. And if an interpretation can be understood without more ado, then it is false that every utterance needs an interpretation. If that is not so, then we are launched upon an infinite regress. (Hacker:1997, 301)

This last point echoes Wittgenstein's claim that there must be a way of grasping a rule that does not involve interpretation.[33] However, as a criticism it is out of place and betrays a misunderstanding of Davidson's argument. Hacker's diagnosis of Davidson's error shows why. He writes: "Davidson's idea that all understanding requires interpretation rests ultimately upon the error of supposing that what we hear when we hear the speech of others are mere sound patterns" (ibid., 303).[34] This, he argues, misrepresents understanding, which, as he sees it, is akin to visual perception – just as we don't see mere patches of colour, from which we infer the object before us, but coloured

---

[32] Hacker takes the definition of interpretation as the substitution of one expression for another. According to Davidson, this accurately describes translation, not interpretation. As we shall see, their difference in this regard can be traced to the difference in their respective understandings of what constitutes successful communication.

[33] Wittgenstein, §201.

[34] In fact, Davidson explicitly rejects this idea. For example, in "The Myth of the Subjective" he says that "the idea that there is a basic division between un-interpreted experience and an organizing conceptual scheme is a deep mistake, born of the essentially incoherent picture of the mind as passive but critical spectator of an inner show" (Davidson: 2001d, 52).

objects themselves, neither do we infer the meaning of an utterance from mere sounds heard. Rather, we perceive the meanings of the words directly. This is possible because speaker and hearer have mastered the same language, which for Hacker means that they have both learned the same set of linguistic rules or conventions that allow them to perceive the meanings that make each other's speech intelligible. It is this "conventionalist" view of meaning underlying his view of linguistic mastery that lies at the bottom of Hacker's rejection of Davidson's position. With that in mind, let's re-examine what Davidson has to say on the matter.

3.2.d Davidson's Denial of the Conventional(ist) View of Communication

Recall Davidson's claim regarding the potential difference in meaning between my and my audience's utterance of 'Wagner died happy' (Davidson: 2001c, 13). If I am asked to elucidate the meaning of the utterance 'Wagner died happy,' I may offer any number of paraphrases or translations; however, in the end I can do it no more accurately (or less fallibly) than by providing the truth conditions for the claim in question: "My utterance of 'Wagner died happy' is true if and only if Wagner died happy" (Davidson: 2001c, 12; 2001e, 66). Now in one way this is obviously tautologous and not very informative; if my hearer did not understand my first pronouncement of 'Wagner died happy' (on the left side of the bi-conditional), she will not understand the second occurrence of the sentence (on the right side of the bi-conditional). However, this criticism, which Hacker levels in his critique, misses the point. If I make the utterance, it

is tautologous. However, this will not necessarily be the case for my hearer (who utters, "His utterance of 'Wagner died happy' is true if and only if Wagner died happy"), for there is no guarantee that the use to which both of us put the words uttered will be the same. As Davidson puts it,

> The speaker, after bending whatever knowledge and craft he can to the task of saying what his words mean, cannot improve on the following sort of statement: 'My utterance of "Wagner died happy" is true if and only if Wagner died happy. An interpreter has no reason to assume this will be his best way of stating the truth conditions of the speaker's utterance. (Davidson: 2001c, 13)

Thus, there is a difference in the two ascriptions, since there is no guarantee that the third-person will mean the same as I by the sentence 'Wagner died happy.'

This is not to deny that there is much overlap in the way interlocutors who are said to speak the same language regularly use the same words and sentences. Davidson's point is only that such overlap is not essential for successful communication. What is necessary is that a speaker speak in such a way that her hearer be able to interpret what she says. But this does not require that interlocutors share a set of linguistic conventions that constitute the source of normativity for correct word usage and thus understanding. Rather, as mentioned earlier, it demands that individual speakers use their words with what their interpreters can perceive as regularity. To quote Davidson:

> If you and I were the only speakers in the world, and you spoke Sherpa while I spoke English, we could understand one another, though each of us followed different 'rules' (regularities). What would matter, of course, is that we should each

> provide the other with something understandable as a language. This is an intention speakers must have; but carrying out this intention, while it may require a degree of what the other must perceive as consistency, does not involve following shared rules or conventions. (Davidson: 2001f, 114)

This view reverses the order of explanation found in the conventionalist theory of meaning and communication presupposed by Hacker, who maintains that it is in virtue of a speaker's grasp of and participation in a set of linguistic customs that make up a given language that her utterances have the meaning that they do. And these customs, while they arise out of communal linguistic practice, exist independently of a particular speaker and her intentions. As Hacker puts it, "what the words of English mean is independent of any individual speaker's intentions"(Hacker: 1997, 299).

As Hacker sees it, in arguing that all understanding involves interpretation, Davidson fails to recognise the role that shared practice, or a shared set of norms, plays in linguistic behaviour. Hacker bases his criticisms of Davidson's claims regarding first-person authority on what he takes to be the Davidson's failure to fully appreciate Wittgenstein's 'meaning is use' doctrine. Hacker takes Wittgenstein to be arguing that mastering a language equals mastering a set of conventions (rules or norms for word use) that are generated out of communal linguistic practice.[35] It is because these speakers master the same set of norms determined by that practice that, generally speaking, understanding is

[35] "What an expression means is given by an explanation of meaning, which is *a standard of correct use*" (Hacker: 1997, 296, italics added). In essence, Hacker's view of linguistic norms is similar to that of Putnam and Burge (among others). Each argues that the norms that govern correct word use have a corporeal origin. Yet, each remains wedded to the idea that it is through the grasp of independently existing norms that a speaker is guided in her use of words.

effortless and immediate and interpretation generally unnecessary. But, more to the point, it is also why interpretation (as Hacker defines it) cannot be the source of understanding – ultimately, it is only by learning the conventions that determine the correct use of a language's words that understanding can occur.

The question is, does Davidson use 'interpretation' in the way Hacker suggests? Davidson recognises that in most communicative exchanges there is much commonality in what each interlocutor takes to be what he calls their established 'theory' regarding what the other is saying. However, he also recognises that there will always be much that is not. Still, in spite of this their hearers can understand speakers. This is because, at bottom, the sharing of conventions is not necessary for successful communication. Rather, a hearer must, in the process of actual communicative exchange, be able to 'devise' a theory of her own that corresponds to what the speaker intends her to understand by his use of words. Davidson summarises this point as follows:

> Meaning, in the special sense in which we are interested when we talk about what a word literally means, gets a life from those situations in which someone intends (or assumes or expects) that his words will be understood in a certain way, and they are...Where understanding matches intent we can, if we please, speak of 'the' meaning; but it is understanding that gives life to meaning, not the other way around. (Davidson: 2005, 120)

Having identified the basic point of disagreement between the two, we can now return to Hacker's charge that Davidson is guilty of making the 'cognitive assumption'.

3.2.e The Cognitive Assumption Revisited

Hacker's claim is that Davidson mistakenly thinks that sincere self-ascriptions of the form 'I believe that *p*' are statements about one's mental state of belief and thus constitute a form of self-knowledge (Hacker: 1997, 292). That, in spite of appearances, they are not is shown by the fact that a speaker's warrant for making the claim that he believes that *p* is the same warrant he will have for *p* itself. That is, the avowal of a belief that *p* commits one to the truth of *p* itself (recall Moran's Transparency Condition). As Hacker sees it, the solution to this pseudo-problem – namely, the merely apparent problem of accounting for the asymmetry between first- and third-person avowals of belief given the (mistaken) assumption that they involve different sorts of cognitive achievement – is to recognise that in such cases there is no semantic content attached to the use of 'I believe that...'(in other words, nothing additional is asserted with the addition of this phrase).

There are two problems with this solution. First of all, the claims ('*p*' and 'I believe that *p*'), which are supposedly identical in content, have different truth conditions. For example, my utterance "I believe that the cat is on the mat" is true if and only if I do believe that the cat is on the mat. However, my utterance "The cat is on the mat" is true if and only if the cat is in fact on the mat. The second problem, which is one of the ones we started out with, is this. There is a chance that a significant number of my statements of '*p*' may be false; however, when I make claims of the form 'I believe that *p*', they are almost always true, and this is something that Hacker's critique, which denies truth-evaluability to our self-ascriptions, does not explain. That said, it is crucial to note that by

denying that '*p*' and 'I believe that *p*' are identical in content, we do not thereby deny Hacker's expressivist thesis that both utterances *express* the same belief.

So does Davidson think that sincere self-ascriptions are forms of self-knowledge in the way Hacker asserts? Certainly Davidson does sometimes write in this way. For example, in 'Knowing One's Own Mind' he states: "It is seldom the case that I need or appeal to evidence or observation in order to find out what I believe; normally I know what I think before I speak or act" (Davidson: 2001b, 15).[36] But is he using 'know' here to connote some sort of propositional knowledge regarding the content of our mental states? We have seen that his basic strategy is this: our self-ascriptions are expressed in the form of propositions; therefore, if we can show that we enjoy unique authority with respect to the speech we use in making these utterances, we will have accounted for most, if not all, we need to know about the authority we enjoy with respect to knowledge of our beliefs. First-person authority in speech is therefore explained through a consideration of the requirements of interpretability, as were outlined earlier. But it can be noted that nowhere in that discussion does Davidson mention any special cognitive ability particular to knowledge of one's own mental states, other than perhaps the linguistic know-how that all competent speakers enjoy. If anything, his position runs counter to such a view, for as he stipulates on a number of occasions, the asymmetry between speaker and hearer

[36] Davidson: 2001b, 15. Incidentally, Hacker also quotes this passage to point out what he takes to be a contradiction in Davidson's so-called transcendental argument: On the one hand, Davidson argues that because we know the meanings of our words, we also know our beliefs. But, on the other hand, as this passage indicates, we also know what we believe before we speak or act. However, one could only take this as an indication of inconsistency if one failed to take into account the context in which the passage appears, namely a discussion of the fact that we do not normally come to know our own beliefs through observation of our behaviour.

authority derives from the fact that the latter, but not the former, must rely on interpretive inference (in the "radical" sense) to grasp the meaning of her words (Davidson: 2001c, 12-13; 2001b, 37; 2001e, 66). In other words, the hearer, not the speaker who utters a self-ascription, must do the detective work.

Contrary to Hacker's assertions, then, I think it is evident that Davidson is not guilty of making any sort of 'cognitive assumption'. Hacker lays this charge because he interprets him as arguing that, since they reflect something we 'know,' self-ascriptions are assertions about one's mental states. Such a position goes against what Hacker thinks is the non-assertoric and therefore non-epistemic status of such first-person psychological locutions.[37] This is based in part on his view that because a speaker's warrant for '*p*' is the same as it is for 'I believe that *p*', the utterances amount to the same claim. But, as evidenced by the fact that these locutions have different truth conditions, this does not follow.

Thus, Hacker has confused two ideas, neither of which Davidson has any reason to reject, namely that:

(1) 'I believe that *p*' and '*p*' express the same belief (the key expressivist claim),

and

(2) all the evidence one typically has for the utterance of 'I believe that *p*' will be all that one has for '*p*' (Moran's Transparency Condition),

with the idea that

---

[37] Here, Hacker follows Wittgenstein, who writes: "I can know what someone else is thinking, not what I am thinking. It is correct to say 'I know what you are thinking', and wrong to say 'I know what I am thinking'. (A whole cloud of philosophy condensed in a drop of grammar.)" (Wittgenstein, 222).

(3) 'I believe that *p*' and '*p*' say or *mean* the same thing, which Davidson denies.

But this also points to a lacuna in Davidson's account of first-person authority. Speaker's authority is grounded on the fact that speakers cannot generally misuse their words. The problem is, Davidson does not discuss the distinctive role self-ascribing locutions such as 'I believe that...' play in the authority that is claimed to attach to these utterances. This is needed, since, as his theory is stated, I am just as authoritative regarding the meaning of my words in uttering '*p*' as I am in uttering the self-ascriptions 'I believe that *p*'. But what needs explanation, it has been argued, is why my utterance of the latter, and not the former, is almost invariably true. As it stands, this is not explained.

A more nuanced expressivist reading of self-ascriptions than Hacker provides – one that reconciles (1) and (2) with the denial of (3) -- explains how to fill the gap in Davidson's account. This will be addressed in the following chapter.

3.3 Conclusion

Moran argues that discussions of self-knowledge should move beyond questions of epistemic access, of how we get our pronouncements about our first-order mental states reliably right. This is because he thinks that explaining 'privileged access' alone doesn't touch on the significance of the first-person perspective. Rather, we need to consider the role self-knowledge plays for the subject as rational agent. As he sees it, we are fully rational when we are capable of "genuine" self-knowledge, or self-knowledge gained from the first-person perspective. Still, that he wants to widen the discussion from what

he sees as too narrow a focus on epistemic access assumes, as do the other philosophers considered before him, that the explanation of the security of avowals is to be given in terms of the correctness of our second-order judgments about our first order states (thus the substantive epistemic achievement). However, given that self-ascriptions expressive of genuine self-knowledge are defined by a commitment to those self-ascriptions being grounded upon first-order reasons for the states self-ascribed, it becomes difficult to see what role second-order judgement and belief might play here. Thus, I suggested that Moran's account of the sort of substantial cognitive achievement underlying so-called genuine self-knowledge claims actually points to a deflationary understanding of the authoritative self-ascriptions by which it is thought to be expressed.

This brought me to a consideration of Davidson's deflationary account and Hacker's expressivist critique of it. I argued that Hacker's main criticism – that Davidson is guilty of the "cognitive assumption" (that self-ascriptions normally express knowledge claims about one's mental states) – was based on a fundamental misunderstanding of the idea that the way the speaker uses her words determines their meaning. However, I have argued that even though *Hacker's* expressivist critique misses the mark, Davidson's explanation of authoritative self-ascription in terms of the necessity of semantic authority for speech does remain in need of additional support of a more sophisticated expressivism than Hacker provides. This is the first topic of discussion in Chapter 4.

Chapter 4: Expressivism and Rational Agency

Introduction

In Chapter 3 I argued that Davidson's account of first-person authority can serve as the basis for a non-epistemic understanding of authoritative self-ascriptions, but finished by suggesting that it is in need of the sort of filling out that a properly conceived expressivism about self-ascriptions can provide. Thus, in the first part of this chapter I offer a description of such an account, and consider various objections that have been offered to it. One set of criticisms – namely that of Dorit Bar-On – comes in for more attention than the others. Bar-On offers what she calls a "neo-expressivist" alternative to expressivism that is motivated in part by a desire to accommodate what she takes to be valuable insights of the supervisory model discussed in Chapter 2. I argue that, not only are the adjustments to expressivism needed to accommodate the supervisory model unnecessary, but they undermine her explanation of first-person authority as well. This is because – as I go on to argue in the second part of the chapter – the supervisory model of rationality is untenable; we cannot, nor need we, exercise the sort of higher-order rational control over our mental states that this model presupposes. But what, then, becomes of rationality? Does the fact that we lack the kind of control supposed by the supervisory model somehow undermine our status as rational beings? To assuage such potential worries, and offer a starting point for how we might consider rationality without self-

knowledge, I take a brief look at one version of what I call a "bottom-up" view of rationality, namely that of Donald Davidson.

4.1 An Expressivist Account of Self-Knowledge

4.1.a The Truth-Evaluability of Expressive Self-Ascriptions

As we saw in the previous chapter, according to Hacker, Davidson mistakenly thinks that understanding the utterances of another is a matter of interpretation, and that the asymmetries between self- and other-ascriptions are explained by the fact that the hearer, but not the speaker, must rely on her interpretative abilities to know what a speaker's words mean and thus the thoughts she expresses through her sincere use of them. Hacker disagrees, arguing that interpretation presupposes understanding. As he sees it, Davidson's mistake leads him to the mistaken 'cognitive assumption', or the idea that self-ascriptions are second-order assertions about our first-order mental states, and thus constitute reports or descriptions of them. As we saw, Hacker rejects Davidson's claim that "a person normally knows what he or she believes," not because Hacker thinks that we are especially fallible in this regard, but because he denies that self-ascriptions constitute second-order assertions about one's mental states. Rather, they are expressions of the first-order states they only appear to describe. On Hacker's view, it follows that no semantic content attaches to the 'I believe' (or desire, intend, etc.) clause that makes it seem that such utterances count as reports. Put another way, if, contrary to appearances,

such utterances are not second-order assertions about our mental states, and thus do not constitute reports of them, it follows that they are not truth-evaluable. In saying 'I desire a drink' I express my first-order desire for a drink. Because this utterance only expresses my desire, as opposed to a belief about it, it is neither true nor false.

As alluded to in Section 3.2.e, the problem with this is that self-ascriptions certainly seem truth-evaluable. As Jacobsen remarks in "Wittgenstein on Self-knowledge and Self-Expression", self-ascription of mental states

> bears all the surface marks of truth-evaluable discourse.... [A]n argument such as 'If I desire water, then I will drink; I desire water; so I will drink' appears to be an instance of a familiar valid argument schema. But a valid argument schema is one that carries truth from premises to conclusions. Similarly, 'I do not desire water' certainly appears to be the negation of 'I desire water'; but negation just is the operation which reverses truth values. Only by assigning truth-values to self-ascriptions do we have within easy reach a satisfying account of the surface features of psychological discourse, and so of our normal practice of ascription and argumentation. (Jacobsen: 1996, 19)

As Jacobsen notes, Hacker acknowledges the syntactic fitness of self-ascriptions for truth-evaluability, but still maintains they are not so assessable on the grounds that they are not second-order assertions (ibid., 20-21). We must, Hacker holds, remember their expressive role: it would be just as mistaken to hold an utterance of "My stomach hurts" or "I am scared!" as true or false as it would be a groan of pain or a cry of fear (Hacker: 1986, 298). Suppose this is so. In that case, since Hacker acknowledges that self-

ascriptions are syntactically fit for truth-evaluability, but are not truth-evaluable, the expressivist would seem to owe "an alternative (non truth-functional) account of what only *appear* to be our standard logical operators and connectives, conditional sentences or valid arguments" (ibid.).[38]

Perhaps fortunately, such an exercise may not be needed. The challenge is to see how the truth-assessability of self-ascriptions need not entail second-order assertoric status. We have an opening toward an understanding of how this may be so if we hold to a minimalist conception of truth-evaluability. On that conception, a meaningful sentence of a language counts as truth-assessable if (1) it can occur without (surface) syntactic incongruity as an antecedent-clause in the disquotational schema ('"p" is true if and only if p'), and (2) it has a significant negation.[39] On the grounds of (1) and (2) it seems undeniable that self-ascriptions are truth-evaluable. But, if so, do they not then count as second-order assertions about, and not first-order expressions of our mental states after all? Not necessarily, for truth-evaluability is not a sufficient condition for attributing assertoric status to a sentence – there are many instances of sentences that when spoken satisfy the requirements of the disquotational schema for truth-evaluability and yet are not assertions. For example, on the basis of (1) and (2) an utterance of 'I promise that *p*' counts as truth-evaluable. However, when I sincerely utter 'I promise that *p*', I do not *assert* that I promise that *p*, but promise that *p*. So not only is the utterance truth-evaluable, but whenever adding the performative prefix 'I promise' constitutes the

[38] Simon Blackburn undertakes such a project in "Wittgenstein's Anti-Realism".
[39] See Wright (1992), especially Chapter 1, for a discussion of a minimalist understanding of truth and truth-evaluability.

utterance as a promise, it guarantees that it will be true.[40] Thus, speakers appear to have something like first-person authority with respect to their own illocutionary performances.

4.1.b The Non-Assertoric Status of Self-Ascriptions

An assertion is a truth-evaluable speech-act expressing a belief the content of which coincides with the meaning of the utterance. Thus, an assertoric use of "Grass is green" expresses the belief that grass is green. A sentence that admits of truth – that is, takes the syntactic form of an assertion – might nonetheless serve a non-assertoric function (for example, mentioning an utterance of 'I believe that Wagner died happy' while trying to illustrate a point about Davidson's view of first-person authority). As Jacobsen points out,

> In such cases, the self-ascription is agreed to be divested of any assertoric status it might otherwise have had, for the mundane reason that overt features of context signal that speakers do not, by uttering the sentence, represent themselves as having a belief with the content of the utterance. (Jacobsen: 1996, 24)

What the expressivist account requires is an etiolating or 'divesting' feature that is normally present in the utterance of self-ascriptions. The obvious candidate from an expressivist point of view is the use to which the utterance is put, what mental state it is employed to express. In other words, the expressivist can argue for what Jacobsen terms

[40] See Sinnott-Armstrong (1994).

Expressive Exclusivity, or the idea that a token utterance may only express a single state (ibid.). So, if an utterance of 'I desire a drink' is employed to express a desire, it cannot express a belief that one has that state. This is the idea Hacker appeals to in his critique of Davidson when he denies that such utterances have truth-values (perhaps combined with the additional assumption that only expressions of belief are truth-evaluable). Now it may seem, that the expressivist here is trying to pull a fast one that comes very close to begging the question against one who might argue for a cognitivist understanding of self-ascriptions. However, that cannot be the case, since the cognitivist makes use of the same principle to argue *against* the idea that self-ascriptions, when seen as expressions of second-order beliefs about our first-order mental states, could be expressions of the first-order states they attribute. Indeed, without the Expressive Exclusivity principle, one would be free to argue that a self-ascription of, say, a desire that *p* simultaneously expresses the first-order desire it attributes as well as the second-order belief that one has that desire. In fact, I have already argued that this is just the position to which Moran seems to be committed. However, as I mentioned in my discussion of him, this introduces unnecessary complexity into the understanding of self-ascriptions and the explanation of first-person authority and thus may be avoided. So let us assume for now, as do expressivism's critics, that Expressive Exclusivity is true. With that assumption in place, we are left to conclude that the employment to which a self-ascription is put will determine the mental state expressed through its utterance.[41]

[41] As Jacobsen points out, this idea echoes Frege's distinction between meaning and force: An utterance with a univocal meaning – for example, "You're next" – may, depending on the situation, be used to perform a number of different linguistic acts. It may be used to express a

If so, the next question is: what, if anything, might determine whether an utterance of 'I believe (desire, intend, etc) that *p*' typically expresses a second-order belief that one has the state ascribed or the first-order state indicated by the content clause *p*? Jacobsen argues that expressivists can see the meaning of the psychological term contained in the prefix of a self-ascription as playing such a role. With respect to explicit performatives, he notes that "the mere appearance of the performative verb '*V*' in the prefix ('I *V* that …') introducing the content clause divests those utterances of assertoric status, assigns them another performative status and hence another expressive character" (ibid., 26). So, an utterance of 'I promise that *p*' has the default performative status of effecting a promise, and thus expressing an intention to do *p*, rather than an assertion about that intention. But it should also be noticed that this is so only when such utterances are made in the first-person present-tense – utterances of 'He promises that *p*', or 'I promised that *p*' remain assertions.

Given this, the expressivist may now argue that a present-tense psychological self-ascription shares the same feature; that is, its default expressive status is determined by the meaning of the psychological term that appears in its prefix. As Jacobsen summarises it,

> [Expressivists should see such meanings as signalling] the distinctive expressive character of utterances of self-ascriptions, just as explicit performatives have verbs the meaning of which signal their own distinct performative character when in the first-person present-tense. Where explicit performatives ascribe the very acts of

---

belief, a threat, a promise, or a prediction, and so express a number of different mental states (Jacobsen: 1996, 26).

> speech they also perform, self-ascriptions ascribe the very mental states they also express. Each thereby divests itself of what, on grounds of syntax and truth-assessability alone, would have been assertoric force. (Ibid., 27)

On this view, the meaning of a non-observationally derived self-ascription (that is, one not derived from one's interpretation of one's own thoughts or behaviour) and its assertoric force will diverge. However, as we have seen, this is the same exception normally taken to apply to explicit performatives. Take the example of assertion: by a sincere utterance of 'I assert that *p*', a speaker does not assert that she asserts that *p*; rather, she asserts that *p*. Similarly, in sincerely saying 'I promise that *p*', she does not assert that she is promising that *p*, but rather promises that *p* (in facing criticism for having failed to follow through, it wouldn't do for her to protest that she simply had been mistaken).

4.1.c Expressivism and First-Person Authority: the Connection Between Truth and Sincerity

Jacobsen's explanation of first-person authority is now this. If my self-ascriptive utterance of 'I believe that *p*' serves to express my belief that *p* (as opposed to expressing my belief that I have that belief), and if I am sincere in my utterance of the self-ascription (that is, if I have the belief I express), then it follows that my utterance must be true. In uttering 'I believe that Wagner died happy' I ascribe to myself the very belief that my utterance expresses; assuming I am sincere, I will then have the belief I ascribe to myself. This explains why, when I utter sincere self-ascriptions of my mental states, I will always

get them right. And the fact that they are expressions of mental states, and not assertions about them (in other words, knowledge claims derived from some sort of cognitive act of detection, for example some form of introspection or self-observation), explains why we can make them immediately and effortlessly, since to make a sincere self-ascription just is to be in the mental state ascribed.

If correct, the expressivist account outlined above – which will be the version to which I shall refer from now on by the term 'expressivism', unless otherwise stated – provides a deflationary account of first-person authority. However, it is not without its detractors, and in a moment I shall consider some of their objections. But before that, I wish to return to the earlier discussion of Davidson and see how the expressivist view just described fills out the lacuna noted earlier in his account of first-person authority.

4.1.d Davidson and Expressivism

The expressivist analysis of self-ascriptions should not be construed as a wholesale rejection of Davidson's position on first-person authority. In fact, similarities are to be found between Davidson's view and the expressivist take on self-ascriptions. In 'On Saying That' he writes:

> A certain interesting reflexive effect sets in when performatives occur in the first-person present-tense, for then the speaker utters words which if true are made so exclusively by the content and mode of the performance that follows, and the mode

of the performance may well be in part determined by that same performative introduction. (Davidson: 1984, 107)

This is followed by an example that, although it is offered in a different context (the explanation of indirect discourse) makes just the expressive point outlined above, namely that, in the case of utterances whose verbs have a default performative status, utterance meaning will differ from assertoric content. If I utter 'I assert that *p*', I do not assert that I am making an assertion (namely that *p*); rather, I assert that *p*. Indeed, if the former were the case, then all my assertions would necessarily be false, for the simple reason that my assertion that *p* would only be true if I assert that *p*. But on a non-expressivist reading, I would not be asserting that *p*, but asserting that I assert that *p*. Here is Davidson's example: he notes that if I utter 'I assert that Entebbe is equatorial,' the performative verb 'assert' announces, as he puts it, the expressive character of the utterance, and in so doing makes such (sincere) pronouncements "self-fulfilling" (ibid.). Thus, in uttering 'I assert that Entebbe is equatorial', I do not assert that I assert this, but simply assert it.

Given that this is in keeping with the expressivist analysis outlined above, it is reasonable to argue that, even if he did not make use of them, Davidson had at his disposal many, if not all, of the conceptual tools needed to round out his own explanation of first-person authority. In short, with Davidson's theory of semantic authority, we get an explanation of how it is that we cannot misunderstand the words through which we express our mental states. With a properly conceived expressivism we get an explanation of how we are able to effortlessly and authoritatively ascribe to ourselves the whole range of mental states that can be expressed through those words.

In the following sections, I turn to a consideration of several objections to the expressivist view of first-person authority just outlined.

4.1.e Objections to Expressivism (I): Moran – Self-Ascriptions Report Mental States

In his critical discussion of expressivism Moran rejects the idea that self-ascriptions never serve as assertions by which a speaker reports or describes her mental states. He acknowledges that there are a number of cases where psychological verbs such as 'believe' are not employed to report a state of belief, as when one hesitantly says, "I believe it may rain today." However, he contends that it is implausible to suppose that, after having taken into account these various other uses "it will turn out that *no* such first-person psychological statements actually have a reporting function, that none of them count as something said with the intention of telling another person my thoughts, beliefs and feelings" (Moran, 71). Moran thinks that philosophers are led to this view through a mis-reading of Wittgenstein, whom he takes to reject expressivism. Moran writes:

> It is this type of view [that is, expressivism] Wittgenstein is alluding to when, for instance, he concludes one line of thought about Moore's paradox with the advice, "Don't regard a hesitant assertion as an assertion of hesitancy" (*Philosophical Investigations*, p. 192). That is we are to see the hesitancy expressed by the apparent reference to one's belief (as in "I *believe* it's still raining out") as qualifying the assertion about the rain, and not as describing anyone's state of mind. However, to ascribe [expressivism] to Wittgenstein one would have to understand this passage

> and related ones as not just warning against a confusion to which we may be prone, but as claiming that, for instance, hesitancy can *only* apply to assertions and not to persons and their states of mind. This is not what he says; and had it been what he meant, it would have made less sense to warn against a confusing one thing with another than simply to declare that the very idea of an assertion of one's own hesitancy (or doubt or conviction) can only be an illusion. (Ibid., 72)

Indeed, as mentioned previously, as Moran sees it, such a view would entail what he takes to be the preposterous idea that while one could talk and think about the mental states of others, one would be incapable of doing so with respect to one's own mental life.

Two mistakes are made here. First, with regard to the above passage, Moran misunderstands the intent of Wittgenstein's claim. He suggests that to attribute expressivism to Wittgenstein we would have to see him as "not just warning against a confusion to which we may be prone, but as claiming that, for instance, hesitancy can *only* apply to assertions and not to persons and their states of mind." In fact, it is just the opposite. In telling us not to regard a hesitant assertion as an assertion of hesitancy Wittgenstein is making the expressivist claim that one is not asserting that one is hesitant, but rather is expressing one's hesitancy about the matter in question. In other words, he is claiming that the very idea of an assertion of one's own hesitancy (or doubt or conviction) not arrived at through some sort of self-observation *is* an illusion fostered by the grammar of the utterance. So, an utterance of "I believe it's still raining out" is a

hesitant assertion – one expressing hesitancy -- not an assertion to the effect that I am hesitant.

Now to the second error: As Moran sees it, if expressivism were true, we would be precluded from talking or thinking about and reporting on our own mental lives. The objection assumes that for a self-ascription to be truth-evaluable, and thus serve to report one's mental state, it must count as an expression of a second-order belief (and so as an assertion) about that state. However, as we have seen, if the version of expressivism outlined above is correct, then the truth-evaluability – and thus reporting status – of self-ascriptions is compatible with their nonassertoric expressive character. Put another way, one may report on one's mental state by expressing it in a self-ascription. Indeed, what better way could there be to tell another what one thinks, believes, or feels than by directly expressing it to them in such a form? But if we can communicate our states to one another in such a way, then Moran's objection is unfounded.

4.1.f Objections to Expressivism (II): Wright's Secret Agent Man

Unlike Moran, Wright's criticisms do not turn on a misunderstanding of the basic expressivist position outlined above. His primary objection centres on the (supposed) explanatory work done in the expressivist account by what he calls the appeal to illocutionary distinctions (Wright: 2001e, 358-364). The objection comes in two parts. We are to imagine a secret agent – call him Max – who in virtue of his training is capable of showing no ordinary behavioural signs of pain when tortured. However, his

tormentors, having the requisite instruments and knowledge of the characteristic signs of pain (for example, patterns on the electro-encephalograph, raised heart rate, activation of reflexes in the eye), are able to detect his state of being in pain. It follows that, in terms of the acquisition of knowledge "strictly so conceived" (that is, in terms of some form of justified true belief gained from evidence) of his mental state is concerned, the tortured subject is in no more of a privileged position than his tormentors (ibid., 363). In fact, in the case imagined, the subject is at a distinct disadvantage, for lacking any ordinary behavioural evidence, (1) in his agony he may not be able to attend to the minute but telltale involuntary "outward" signs that would give away his mental state, and (2) he would lack access to the instruments that indicate to the torturers that their techniques are having the desired effects. So, "when it comes down to knowledge, it looks as thought the expressivist account must represent the victim as actually at a *disadvantage*. And that's evident nonsense" (ibid., 363).

The objection seems to be that, because he lacks evidence of his pain, Max is less capable of knowing his state than his tormentors. Wright takes this as nonsensical, but is it? If knowledge, "strictly conceived" is something that requires evidence of some sort, and Max lacks the evidence others have, then yes, he is at a disadvantage. However, it does not follow from this that Max is any less "aware" or "conscious"[42] of his current state (unless, perhaps he is delirious from the pain). Given that Max is *in* the state that his

[42] 'Aware,' 'conscious,' and similar terms must be used advisedly; it must be remembered that with them one is not suggesting that one arrives at such awareness through some sort of judgement.

tormentors can only know through the interpretation of evidence, his lack of knowledge places him at no disadvantage in this regard.

A related objection regards the explanation of our authority with respect to unexpressed (that is, unverbalised or merely thought) self-ascriptions. Here's Wright:

> You may sit reading and think to yourself 'My headache has gone' without giving any outward signs at all. And anyone versed in ordinary psychology will accept that *if* you have that thought, not by way of merely entertaining it but as something you endorse, then you will be right (Authority); that there is no way your headache could have passed unless you are willing to endorse such a thought (Transparency); and that your willingness to endorse it will not be the product of inference or independently formulable grounds (Groundlessness). Thus analogues of each of the corresponding marks of avowals that pose our problem engage the corresponding unarticulated thoughts. It must follow that the correct explanation of the possession of them by avowals cannot have anything to do with illocutionary distinctions. (Wright: 2001e, 364)

The problem is that self-ascriptions merely in thought share all the characteristic features of their more extroverted brethren, the avowals. One strategy for overcoming this objection might be to argue that in fact our unarticulated self-ascriptions do count as expressions of the state self-ascribed after all – they are just of the "unvoiced" variety. This is the line of defence offered by Dorit Bar-On and Douglas Long, who also argue for a variant of the expressivist position. With regard to what they call "avowals proper" they write: "These are sincere, spontaneously volunteered, unreflective utterances (voiced or

silent), such as, "I've got a terrible headache!" or "I'm so frightened of it!" or "I think she's going to fall!"" (Bar-On and Long, 326). It is difficult to make sense of such a "silent" utterance. Happily, such an approach is not needed; as Jacobsen argues, there is a straightforward explanation of how features unique to expressed self-ascriptions can accommodate Wright's concern.

Two uncontroversial assumptions are needed for the explanation. As Jacobsen puts it,

(1) If *what we say* is true, and we think or believe what we say, then *what we think* or believe is true.

(2) If what we think or believe is true, then it will be true *whether or not we say it.* (Jacobsen: 2007, 14-15)

From (1) and (2) it follows that "[i]f what we say is true, and we believe what we say, then what we believe will be true whether or not we say it" (ibid., 15). So, if our sincere utterances of self-ascriptions are normally true, as expressivism says they must be, and we believe (hold true) what we say (as we must, given sincerity), then it follows that our sincere self-ascriptive thoughts must also be true, whether or not they are ever uttered. So Wright's argument doesn't undermine the expressivist account.

4.1.g Objections to Expressivism (III): Heal – Sincerity Without Truth

In "First-Person Authority," Jane Heal argues that expressivism fails as an explanation of first-person authority on the grounds that sincerity of utterance does not guarantee the truth of self-ascription. She sets up her argument with an analogy by distinguishing

between "natural" and "personal" promise. As Heal describes it, personal promising refers to what we normally talk about when discussing or using the term 'promise' (for example, by A's uttering 'I promise to take you to the park,' to B, A commits himself to taking B to the park). According to the expressivist explanation of first-person authority, personal promises provide a good analogue to self-ascriptions – "I promise that *p*" ascribes a promise to the speaker, the truth of which is guaranteed whenever that utterance constitutes her as making a promise (recall the earlier claim in 4.1.a that "speakers appear to have something like first-person authority with respect to their own illocutionary performances"). Natural promise refers to what we "show" when we evidence a disposition to do or be capable of something. As Heal puts it, "A naturally promises to *p iff* A shows a real tendency to *p* and so entitles an observer to expect that A will *p*" (Heal, 277). To take Heal's example, little Sandra may show natural promise as a mathematician because she has mastered algebra at a very young age. Natural promise constitutes a good analogue of many mental states, which are sufficiently like dispositions or tendencies. Having outlined these two senses of promise, we are then asked to imagine the "blinkered philosopher," who is only capable of interpreting utterances of 'I promise to *p*' as statements of natural promising. The question then is: could the blinkered philosopher make sense of the authority that people seem to grant to self-ascriptions of personal promise from the expressivist point of view?

The expressivist, who is Heal's blinkered philosopher, thinks he can explain a speaker's authority concerning her natural promise (in other words, her mental states) from the authority she has regarding her personal promises (that is, her illocutionary

acts). So what, according to Heal, would an expressivist account of natural promising (that is, 'normal' promising from the blinkered philosopher's perspective) look like? Speakers would have to be trained such that they would be disposed to utter 'I promise to *p*' when they do show such (natural) promise; such an utterance would be an expression of their promise. Heal remarks that this may look somewhat bizarre, but that it is nonetheless imaginable. At any rate, she goes on, we need not bother with this issue, for it is not where she thinks the major problem lies (we will get to what she takes that to be presently). However, it is a greater difficulty than she recognises, at least for her understanding of expressivism. Take Sandra's utterance of 'I promise to be a fine mathematician'. As Heal has it, Sandra comes to this disposition to make this utterance after being trained to do so in association with her natural promise to be a fine mathematician. However, such an utterance would not be an expression of that promise; rather, solving a math problem would be an expression (and evidence) of the developing disposition to be good at math. On the expressivist view, an utterance of 'I promise to be a fine mathematician' would be an expression of the belief that one has the disposition, and not, as she thinks, an expression of the disposition itself.

Moving on to what she takes to be the major problem, Heal argues that, contra expressivism, sincerity does not guarantee the truth of a self-ascription (ibid., 280). Recall that, according to expressivism, the link between truth and sincerity is crucial to understanding first-person authority. On the expressivist view, a sincere utterance of a self-ascription cannot normally be false, since the state reported just is the state expressed. But if we allow that an utterance may be sincere "provided merely that it is

produced spontaneously and in good faith," then sincerity does not guarantee authority (ibid.). As she sees it, nothing in the training that sets up the disposition to utter the self-ascription of natural promise "shows that it must be so effective that no putative expression ever occurs spontaneously and in good faith but in the absence of the state to which the training designs it to manifest" (ibid.). Or, as she otherwise puts it, nothing in the training rules out the possibility of sincere "false positives". And if this is so, then sincerity cannot play a role in the explanation of the authority. And, she adds, the same problem will apply to psychological self-ascriptions as well.

There are two related points of interest here. The first is not directly related to Heal's objection, but it is a problem nonetheless. One might ask: if, as Heal supposes, the disposition to utter the self-ascription is inculcated in the subject in association with another disposition (that is, a natural promise to *p*), how could she sincerely yet falsely make such a self-ascription? In other words, how could that disposition (natural promise) ever be absent, which would make the utterance false? This could only be so if the self-ascription of natural promise were taken as a second-order assertion about the state and not the expression of it. This takes us to the second issue. As both the cognitivist and expressivist agree, for an utterance to be sincere the speaker must have the state she expresses; where they disagree is about what that state is (a first-order state, or a second-order belief about it). For a false self-ascription to be sincere, the state the speaker expresses must be a second-order belief that she has the (absent) disposition to show natural promise. But if the criticism of expressivism rests on this possibility, then it begs the question against it.

4.1.h Objections to Expressivism (IV): Bar-On and Epistemic Expressivism

Dorit Bar-On (2004) argues for an understanding of our authoritative self-ascriptions that, while sharing certain ideas with the expressivist view argued for above, differs from it in several significant respects. Her "neo-expressivist" (as she terms it) account accepts the core expressivist insight concerning the explanation of the security our authoritative self-ascriptions, namely that they express the states they self-ascribe; however, she rejects what she sees as certain mistaken conclusions that some have drawn in light of it. I shall consider three related criticisms: first, that what I earlier called a more nuanced form of expressivism entails an overly strong infallibility of the first-person; second, that this expressivism entails a form of irrealism regarding the existence of mental states; and third, that contra this version of expressivism, there is a story to be told about how authoritative self-ascriptions count as substantial or "robust" (to use her term) self-knowledge.

The first two criticisms are grounded on misunderstandings of the expressivist argument.[43] Concerning the first, Bar-On points out that, as she puts it, just because a subject's self-ascription expresses a mental state, it need not follow that with it she expresses *her* mental state. For example, on the expressivist view, a subject's utterance of 'I love you' is normally taken to express love. However, if the subject does not in fact love the intended hearer, then in uttering 'I love you' she does not express her love (since

[43] These two criticisms were first articulated in Bar-On and Long, p. 328, n. 32 and p. 333, n.37.

she has none to express). This distinction, she suggests, is easy to miss, because "'express', like 'show', as well as verbs of perception and some epistemic verbs, is a success verb. If we say that a subject has expressed her anger, or joy, we imply that the subject is indeed in the state cited" (Bar-On, 280). Now the criticism of expressivism seems to be this. Because the use of 'express' normally implies that one is in the state one expresses – that in avowing a state one is in the first-order state self-ascribed – expressivism is led to suppose that all expressions of states in self-ascriptions must be accompanied by the state self-ascribed. Hence, Bar-On claims, expressivism mistakenly sees our avowals as "strongly infallible" (ibid., n.43).[44]

If I have understood her correctly, there is some merit to what she says, but not as an objection. That is, there *is* a sense in which our avowals are "strongly" infallible, namely when they are sincere. Indeed, the link between truth and sincerity is what accounts for first-person authority. However, nowhere is it suggested that sincerity in avowal is guaranteed. Unfortunately, Bar-On seems to have overlooked this qualification, at least if she thinks that expressivism has missed the distinction between expressing a state and expressing one's state – which is just the difference between being sincere and insincere[45] – and thus holds that there are no cases where expressive self-ascriptive utterances (viz., those not arrived at through self-observation) may be false.

The second criticism is related to the first in that it too concerns the connection between truth and sincerity in authoritative self-ascription. The critical claim is that expressivism entails that "there simply are no independently existing mental states of

[44] See also see p. 317, n. 24 and p. 325, n. 28 where the claim is repeated.
[45] See the discussion of Heal in 4.1.g.

subjects to ground the cognitive success *or* failure of mental self-ascriptions" (ibid., 353).[46] This assertion is made within the context of a discussion of Elizabeth Fricker's analysis of what might be meant by the idea that self-ascriptions lack "cognitive achievement." One way of understanding this is to suppose that "[it] is not the case that a person's first-level mental states and her judgements self-ascribing them are ontologically distinct states, and that the second reliably track the first" (Fricker, 173). Given that, according to expressivism, the state self-ascribed is the state expressed, this is something to which expressivism is committed. But does it follow, as Bar-On seems to think, that for expressivism there are no independently existing mental states of affairs – that is, states that exist independently of particular expressions of them – that can ground the cognitive success or failure of self-ascriptions (that is, their truth-evaluability)?

Bar-On supposes that the expressivist may be led to this conclusion through a consideration of ethical expressivism. According to Bar-On, the ethical expressivist offers the expressivist view,

> not only as a positive claim about the expressive force or function of ethical utterances, [namely that they express feelings and attitudes, and thus are not assertions abut ethical matters of fact,] but also as a "non-cognitivist" negative claim to the effect that there are no ethical properties for ethical statements to refer to. (Bar-On, 305)

According to Bar-On's ethical expressivist, ethical utterances express pro- or con-attitudes, and not assertions about independent ethical states of affairs. From this it is

[46] See also Bar-On and Long, p. 332.

inferred that there are no ethical properties to which ethical statements might refer. As Bar-On sees it, the expressivist, wanting to preserve ethical non-cognitivism, imports the same sort of move into his understanding of authoritative self-ascriptions (Bar-On and Long, p. 333, n. 37; Bar-On, p. 353, n.10). Because he assumes that authoritative self-ascriptions express first-order states as opposed to second-order assertions about those states, the expressivist is led to conclude that there are no mental states to which those self-ascriptions may refer and which may ground their truth-evaluability (their "cognitive success or failure").

This may seem a curious charge, given that, as Bar-On herself notes, the expressivist argues for the truth-evaluability of authoritative self-ascriptive discourse (Bar-On, 317, n. 24). It may be that she thinks that the way he argues for this – by arguing for a minimalist conception of truth-evaluability, where 'I desire a drink' is true if and only if I desire a drink – somehow divorces truth-evaluability from the idea that there are independent states of affairs that figure in the determination of the truth of an utterance (what Bar-On calls the "ontological denial" [ibid., 354]). But it is difficult to see how a minimalist conception of truth-evaluability, where 'I desire a drink' is true if and only if I desire a drink, entails such an idea. At any rate, to make such a claim one must ignore the connection between truth and sincerity contained in expressivism's account of first-person authority. For, on that view,

> an utterance *reports* that mental state the presence (or absence) of which makes the utterance true (or false); it *expresses* that mental state the presence (or absence) of which makes the utterance sincere (or insincere). The dispute between cognitivism

and expressivism has concerned whether our self-ascriptions report or express our thoughts and feelings. On the view recommended, a single utterance may both report and express the same (token) mental state. (Jacobsen: 1996, 30).

If so, then in spite of the fact that such self-ascriptions do not serve as second-order assertions, there is no reason to suppose they entail any sort of "ontological denial" as Bar-On conceives it. According to expressivism, the state the speaker expresses in reporting her first-order mental state is the very first-order state she ascribes to herself in that report. In other words, in a sincere self-ascription the expressed state and the state reported are the very same state, and thus they do lack ontological distinctness. However, in no way does this denial of ontological distinctness entail that the state expressed cannot exist independently of its expression.

This brings us to the third criticism. According to expressivism, since their expressive character explains the security of sincere self-ascriptions, there is neither need nor room for any sort of substantial epistemological account of the authority subjects enjoy with respect to such self-ascriptions. Due in part to reasons just discussed, Bar-On disagrees. As she sees it, we need not suppose that because a self-ascription expresses the subject's first-order state that it is therefore barred from simultaneously expressing her second-order belief that she has it (Bar-On, 307-310). And, if we agree that self-ascriptions can express such second-order beliefs, then, contra expressivism, the possibility of "robust knowledge" of one's mental states opens up. The path to this is as follows. First, explain why it is reasonable to suppose that self-ascriptions may have a "dual-expressivist" character, in other words express both the first-order state self-ascribed as well as the

second-order belief that one has that state. Second, explain the way in which self-ascriptions may count as second-order beliefs that is in keeping with the unique expressive character of self-ascriptions. Third, show how such beliefs may count as "robust" knowledge.

Recall that, in the debate between cognitivism and expressivism as it was described above (see 4.1a, 134-135), both were agreed on the idea of expressive exclusivity. Bar-On thinks we ought not feel beholden to this principle because, while the expressivist explanation of security of self-ascriptions is a cogent one that should not be abandoned, opting for dual-expressivism makes it easier to render plausible what most philosophers take to be an obvious fact, namely that self-ascriptions express self-*knowledge*, and so justified true belief. Bar-On thus appears to want to have her cake and eat it too. That is, she appears to offer the dual-expressivist thesis as a way to keep the explanatory benefits of the expressivist account of the security of self-ascriptions without taking on board what appear to be its other counter-intuitive consequences (chief among them its "startling deflationary view of self-knowledge" [Bar-On, 353]).

Suppose we keep open the possibility of dual expressivism. We are then faced with the matter of how we might conceive of the requisite second-order belief expressed by a self-ascription such that it is consistent with that self-ascription's first-order expressive character. The problem, as Bar-On sees it, is that *qua* expressions of first-order states, self-ascriptions do not represent self-judgements – how we normally get our mental states right does not involve any sort of (what she calls) "recognitional judgement" (ibid., 170). So, even if we could plausibly argue that self-ascriptions express second-order beliefs,

the difficult question would remain of how such beliefs could stand in the right epistemic relation to the facts they articulate to count as knowledge (let alone privileged knowledge). Bar-On proposes that the matter be resolved by distinguishing between two senses in which one may believe that *p*: "In what we may call the *opining* sense, one believes that *p* if one has entertained the thought that *p* and has formed the active judgment that *p* on some basis, where one has (and could offer) specific evidence or reasons for the judgment" (ibid., 363). As she notes, the Neo-Expressivist may wish to deny that self-ascriptions express such beliefs – for one thing, as per the asymmetries between self-and other-ascriptions, it is normally deemed inappropriate to ask of a speaker *how* she knows that she believes that *p*. However, she says, "there is a second, more liberal sense of belief, in which a subject believes that *p*, provided (roughly) that she would accept *p* upon considering it" (ibid.).

The suggestion is that we can make use of this second sense of 'believe' to show how self-ascriptions express second-order beliefs that count as substantial knowledge claims about our first-order mental states. First, with respect to belief, Bar-On writes:

> Let us assume (with the Neo-Expressivist) that [an] avowal need not represent a belief I have acquired or a judgement I have formed regarding my present state on this or that basis. Even so, the avowal's product semantically expresses that self-judgement. It represents a proposition concerning a present state of mine, which, we may assume, I understand perfectly well. This proposition can reasonably be regarded as something I hold to be true, where, to repeat, holding true requires only that I would accept *p* if I were to consider it. If so, then even in avowing "I am

feeling thirsty" I am not opining that I am feeling thirsty, I may still be said to believe it, at least in one sense of belief. (Ibid., 365)

But, she adds, we can say something stronger than this. The second-order belief expressed in a self-ascription is not one that we merely hold true, in the minimal sense that we would assent to *p* if queried on whether *p*. Rather, given that self-ascriptions are intentionally offered, the second-order beliefs they express will be intentionally self-ascribed. It is argued that this makes for a more robust, "self-ascriptive" sense of belief than mere holding true, where "[s]ubjects can be credited with the relevant [second-order] beliefs to the extent that they can be seen as intentionally issuing self-ascriptions that represent those beliefs when avowing" (ibid.).

With a suitable notion of self-ascriptive second-order beliefs, as well as an explanation of their security, in place we may now inquire into the sense in which such beliefs may be considered justified and thus constitute substantial knowledge (that is, knowledge that involves some sort of cognitive achievement). Bar-On notes that at first sight this may seem a difficult task, given that, according neo-expressivism, our self-judgements are not grounded on any "distinct epistemic basis"; that is, there is nothing the subject does – for example, some sort of "recognitional judgement" – or need know that underlies the second-order belief and which may provide justificatory reasons for it. However, similar to Burge, she suggests that we may "relax" our reading of the notion of justification to include a type of warrant that does not invoke the usual epistemic capacities. With this in mind, she surveys three different ways an account of justification might be pursued that satisfy this criterion: what she terms the low, high, and middle

roads to self-knowledge. Each of these is found wanting in some respect, which leads her to offer a synthesis of the three that, while not a final explanation of what she takes justification to be, may serve as a basic outline of the form such an account could take.

The low road to knowledge takes a causal/reliabilist form (ibid., 369-373). On this view, the subject's second-order beliefs are justified to the extent that the self-ascriptions by which the subject's second-order beliefs are expressed are reliably caused by the appropriate underlying mental states. According to neo-expressivism, utterances of, say, 'I believe that *p*' express both my first-order belief that *p* and my second-order belief that I have such a belief. To the extent that such utterances are reliably caused by the first-order beliefs they express (that is, to the extent that I am reliably sincere in my utterances), then on the reliabilist account the second-order beliefs expressed by such self-ascriptions will be reliably and non-accidentally true and thus justified. As such, they will count as knowledge. While this view is somewhat congenial to the neo-expressivist understanding of the dual expressive character of self-ascriptions, it does, as Bar-On sees it, have its drawbacks (ibid., 371-373). Chief among them is that it cannot accommodate what she calls "the commonsense belief" that self-knowledge is in some sense a privileged form of knowledge. On the neo-expressivist-reliabilist view, there is nothing unique about self-ascriptions *qua* self-knowledge – in this respect they are just like any belief that arises from some sort of reliable causal mechanism.

In this regard, the "high road" approach makes the opposite claim; according to it, self-knowledge is unlike any other sort in that it is a necessary component of our rational nature (ibid., 373-381). Some who take the high road argue that the capacity for non-

observational self-knowledge is a necessary condition for theoretical and practical deliberation. As already discussed (see 2.2h), Burge offers a detailed account of why this might be thought to be so, and Bar-On offers a neo-expressivist adaptation of his transcendental argument as a summary of one way the high road approach might be construed:

> If successful rational and practical deliberators are possible, deliberators' own avowals of present beliefs, thoughts, preferences, hopes, feelings etc. must enjoy a special epistemic status. These avowals enjoy a special epistemic status only if the deliberator possesses a special entitlement to the judgments expressed by those self-ascriptions. Since we, normal human beings, do successfully engage in rational and practical deliberation, we must possess a special entitlement to the judgments semantically expressed by our avowals. (Ibid., 379)

Bar-On thinks there is much to admire in the high road's connection of self-knowledge to our rational capacities and its emphasis of the extraordinary status of self-knowledge, and in this way it improves on the low road reliabilist account. And, like the low road approach, it is congenial to the neo-expressivist view of self-knowledge in that it offers an account of justification that doesn't involve the kind of "recognitional" judgement usually associated with substantial epistemic achievement. However, she finds the transcendental approach problematic, suggesting that it may have the order of explanation backwards. Rather than see the knowledgeable status of self-ascriptions determined by the role our second-order beliefs play in our rational deliberations, might it not be that we

are capable of such rational deliberations because those beliefs represent something we are in a special position to know?

What she has in mind here seems to be this. Both the high and low road approaches offer a general account of entitlement that makes no reference to the subject's epistemic position in *particular* instances where privileged self-knowledge claims are made. But, she says,

> it may be thought that if I am said to have special knowledge of some of my states that no one else can have, it is partly because of a special relation that I have to the subject matter of my knowledge, or due to my being in a special position to have that knowledge in the relevant circumstances. (Ibid., 381)

The high and low road approaches "de-personalise" self-knowledge, taking it out of the hands of avowing subjects. While each provides a suitably non-recognitional understanding of justification, neither makes room for the idea that successful self-ascription still involves some sort of epistemic achievement on the subject's part that will allow for a robust understanding of privileged self-knowledge.

This brings us to the middle road approach, which places the focus on the act of self-ascription (ibid., 381-388). In normal cases our authoritative self-ascriptions are complex intentional actions involving the mastery of self-ascriptive expressive means through which the subject is able to "speak her mind". Still, as expressions of one's state, such acts are not *underwritten* by any epistemic achievement. As Bar-On puts it, typically when she self-ascriptively expresses her mental state, that state is not an "epistemic target" for her (ibid., 386). The non-recognitional character of the self-ascriptive

judgement means that she is immune to errors in judgement of that kind. This is not to say that false self-ascriptions are impossible; rather, it simply means that they will not be the consequence of any sort of recognitional error. The middle road claim is that the immunity to such error should be seen as providing a sort of default entitlement to one's second-order beliefs expressed by one's self-ascriptive utterances. As Bar-On summarises it,

> A subject whose mental self-pronouncement is taken to be an avowal will be credited with epistemic entitlement by default, so long as we take her to be a normal subject with normal expressive capacities. If her avowal is thought to be true, as it normally will be, then she will be credited with knowledge of her present state of mind. (Ibid., 386)

The middle road approach to self-knowledge takes into account the special position the subject is in with respect to her mental states (only she is in a position to express them). It connects the explanation of self-knowledge to the unique expressive character of self-ascriptions and therefore fits best with the neo-expressivist view of self-ascriptive security. However, as Bar-On sees it, it is also not without its problems. For one thing, it omits any consideration of the relation between the act of self-ascription and the truth of those utterances. Furthermore, it leaves out any mention of the other features of self-knowledge that Bar-On agrees are important for our understanding of it, viz., the role it plays in our status as rational deliberators as emphasised by the high road account (ibid., 388). Consequently, Bar-On offers a fourth possibility, which she describes as a synthesis of the three approaches just reviewed.

While she refrains from providing a detailed account, Bar-On does offer the following as a central component of the approach (ibid., 388-396). The key idea is that, on a neo-expressivist reading of self-ascriptions, the rational cause of the self-ascription, namely the first-order state it expresses, also provides the epistemic reason for it. An utterance of 'I desire that *p*' is rationally caused by the subject's desire for *p*; in some cases – for example, in the case of spontaneous utterances, or self-ascriptions in thought – it may be the only cause.[47] In addition, she suggests, this first-order state may also be an epistemic reason, that is, it may also be what justifies the subject's second-order belief that is expressed by the self-ascription. Reiterating a point made earlier, she notes that "the first-order state is not a justifier in the traditional sense, since it represents no epistemic effort on the subject's part. But," she continues, "the subject is still epistemically warranted – warranted simply through *being* in the state..." (ibid., 390). Furthermore, the connection to self-ascription makes room for a robust account of self-knowledge. Despite the fact that authoritative self-ascriptions do not involve any epistemic effort "the avowal can still be said to represent an epistemic achievement on the subject's part.[48] For the self-ascription semantically expressed is not something that merely pops into the subject's head; it is *epistemically grounded* in the avowed state" (ibid.).

Clearly there is much to consider in the analysis of this view; however, undertaking such a task would take us too far afield. Still a few comments are in order. First, perhaps

[47] This supposes that the thought 'I desire that *p*' can express one's state. I have already mentioned my doubt regarding such a claim.
[48] Albeit an effortless one.

curiously, there is one option with respect to warrant Bar-On does not consider. Accepting that self-ascriptions have a dual-expressive character, we might note that their security is directly tied to our sincerity in uttering them. That is, to the extent that we are sincere, our second-order beliefs will be true. So, it might be concluded, our entitlement to our second-order beliefs is grounded on our status as sincere speakers. Given that one need not know that one is sincere to be so entitled, such an account would be in keeping with the lack of epistemic effort that Bar-On takes to characterise self-knowledge. I just described this as a curious omission; but, it can be noted, it is consistent with her entire discussion of the neo-expressivist picture, in that nowhere does she draw the link just mentioned between truth and sincerity in avowal that underlies the expressivist account of first-person authority.

With respect to the argument she does present, one thing it shows is the complexity and work required to arrive at an epistemic expressivist account. For example, there is what some may see as the unorthodox related ideas of second-order belief as holding true, and "epistemically effortless" epistemic achievement. Other more particular idiosyncrasies may follow. For example, it would seem that we must also postulate a class of self-ascriptive utterances – those that express first-order beliefs – that admit of two radically different epistemologies: an "ordinary" one for our everyday first-order beliefs, and the special one that applies only to self-knowledge. And, with respect to the latter: while at first glance it might seem plausible that a self-ascribed first-order *belief* might justify the second-order belief that one has it (given that most accept that beliefs may be justifiers of other states), it is questionable that other sorts of first-order states (for

example, desires, pains, or hopes) could similarly serve as epistemic grounds for utterances (or thoughts) that attribute them. For this requires holding that states like desires and sensations can – like beliefs – count as reasons for belief. But this is a problematic claim on any but a causal theory of justification (a matter of which Bar-On provides no discussion).[49] While these sort of consequences do not necessarily entail rejection of the view, it is fair to ask: if expressivism provides all we need in terms of an explanation of (non-epistemic) first-person authority, is going to such lengths to preserve the idea that our authoritative self-ascriptions constitute a form of knowledge really necessary?

In response to this, Bar-On might point to the relation, emphasised by "high roaders" such as Burge, Shoemaker, and Bilgrami, between self-knowledge and our status as rational agents, as reason to pursue the neo-expressivist account. For, she argues, it provides a tidy explanation of in what this connection consists:

> On the present proposal, the connection is captured by noting that, when I speak my mind, *I proclaim the very states* – the thoughts, hopes, wants, pains, etc. *p* that *move me in thinking and in acting at the same time as I ascribe those states to myself.* My avowals can be seen as offering up the very states that *provide reasons* for what I think and do, as well as having those states as their own epistemic reasons. (Ibid., 396)

With self-ascriptions we "speak our minds", thus offering up to others the states that motivate our thoughts and actions. What remains unclear, at least from this passage, is

[49] Thanks to Rockney Jacobsen for pointing this out to me.

what role second-order belief might play in the maintenance of rationality, as the supervisory model of the high roaders supposes. For example, Burge argues that second-order belief is necessary for the rational regulation of our first-order states, and it is in virtue of this fact that they count as knowledge.[50] As he and other high roaders (for example, Shoemaker) argue, if we remained unaware of our mental states, we would have no way of engaging in the requisite reflective activities needed to maintain a coherent mental life.

But, as Bar-On agrees, with expressivism we have an explanation of how we are able to realise that awareness without any higher-order epistemic accomplishment. So, even if we agreed that some sort of reliable self-awareness of our mental states were required to maintain a rational coherence (this will be discussed in the following section), the expressivist account of self-ascriptions would seem to obviate the need for the kind of effort undertaken by the neo-expressivist to formulate an account of self-ascriptions as expressive of self-knowledge. Bar-On sees this conclusion as something to be avoided. She does not, as she says, want to be consigned to a deflationary view of self-knowledge (ibid., 388). However, perhaps this may be something to be embraced, especially if the very idea of self-knowledge persists as a remnant of a Platonic/Cartesian view of mind that originates in a misunderstanding of the expressive grammar of self-ascriptive utterances that Bar-On correctly identifies.

[50] In other words, the warrant for our second-order beliefs is explained by this rational necessity. However, if this is so, then Bar-On is incorrect in her criticism that Burge fails to offer an explanation of what renders the second-order belief supposedly expressed by self-ascriptions knowledge.

I have suggested that Bar-On's insistence on an epistemological construal of self-ascriptions introduces unnecessary complexity into her account. But, what is worse, it also creates dangerous tensions within her version of expressivism. Let's recall what distinguishes an expressivist from a cognitivist reading of self-ascriptions. According to cognitivism, all self-ascriptions express beliefs only, and so count as assertions. But the expressivist points out that self-ascriptions may, and typically do, serve different illocutionary purposes, and thus have different expressive functions. That is, an utterance of "I want ice cream" may serve as a request for ice cream, and so express a desire for ice cream. But if it constitutes a request, it isn't an assertion; with my utterance I don't assert that I want ice cream, I ask for some. Showing that not all self-ascriptions equal assertions, and so may express something other than belief, opens up the path to the explanation of first-person authority that expressivism then presents. But Bar-On's insistence on making room for the expression of second-order belief in self-ascription runs counter to this explanation.

4.2 Rationality Without Self-Knowledge

4.2.a A Brief Review

As I have presented them, Shoemaker, Burge, Bilgrami, and Moran all argue for a "non-empirical" (in Boghossian's sense of the term) yet substantial account of self-knowledge that ties the authority thought to accrue to a range of our self-ascriptions to our status as

rational agents.[51] While Shoemaker, Burge, and Bilgrami each has his own distinctive argument, they share an emphasis on the relation between critical reason and rational agency (what I have called the supervisory model of self-knowledge). Moran's view differs in this respect. For him, the connection between self-knowledge and agency is delineated in terms of the special kind of commitment to first-order reasons our second-order beliefs about our mental lives must express if they are to be instances of "genuine" or "first-personal" self-knowledge. This is why I figured him as a transitional figure between those who argue for the supervisory model and the non-epistemic expressivist account offered later. I argued that Shoemaker's, Burge's, Bilgrami's, and Moran's views suffer from difficulties that undermine their plausibility. I then argued that Davidson's account of first-person authority, when supplemented with a properly conceived expressivist understanding of the first-order expressive character of self-ascriptions, provides us with a non-epistemic explanation of the asymmetries that are taken to characterise our sincere self-ascriptions. If so, we may ask what becomes of the idea of an essential connection between self-knowledge and rational agency? In what remains of this chapter I shall examine the general plausibility of such a view and offer a suggestion for what shape a non-supervisory view of rationality might take.

[51] This is not an exhaustive list. For example, Charles Siewart's view is in general agreement with sort of view to which these philosophers subscribe (Siewart: 2003). And, as we just saw, while Bar-On offers an epistemically deflationary account of the security of self-ascriptions, her agreement with many of the connections these authors draw in this regard motivates her argument that our authoritative self-ascriptions nonetheless count as expressions of self-knowledge.

4.2.b Rational Agency and Reflective Control

While the specific orientation of each discussion differs, Shoemaker, Burge, Bilgrami, all draw a connection between our capacity for self-knowledge and rational agency. Each subscribes to what David Owens calls the idea of "reflective control," that we exercise control over our first-order mental states through normative second-order judgement about the probative force of the reasons we have for those states (Owens, 4). In this way we posit our freedom with respect to those states, and are justifiably held accountable for them. We see this idea at work in Shoemaker's claim that deliberation on what to believe and do involves agency, and "that the agency involved in deliberation essentially involves self-knowledge (Shoemaker: 1996a, 28). As he puts it, we are not merely the subjects of our beliefs; through reviewing and deliberating on the quality of the reasons we have for them we decide what it is we ought and shall believe. As we saw earlier, Burge states the idea in more detail:

> As a critical reasoner, one not only reasons, one recognizes one's reason as reasons. One evaluates, checks, weighs, criticizes, supplements one's reasons and reasoning. Clearly, this requires a second-order ability to think about thought contents or propositions and rational relations among them. ... For reasoning to be critical, it must sometimes involve actual awareness and review of reasons; and such a reviewing standpoint must normally be available. ... [T]o be fully a critical reasoner, one must be able to – and sometimes actually – identify, distinguish, evaluate propositions as asserted, denied, hypothesized or merely considered. (Burge: 1998c, 246-247)

It is in virtue of this capacity to review and reasonably confirm and correct attitudes and reasoning by reference to rational standards that we are epistemically responsible (ibid., 258).

Bilgrami also argues that authoritative self-knowledge is tied to the reflective control we exercise over our mental states. Indeed, for him the very idea of first-person authority is a fundamentally normative one, arising as it does out of the justification required for the practices we engage in associated with our holding one another to account for our mental states. He begins with the fact that we generally hold one another responsible for our own mental states, the responsibility of which is predicated on our having reflective control over them. And this, it is argued, presupposes authoritative, rationally necessary self-knowledge – for as Shoemaker and Burge point out, one could not exercise control over one's states through deliberation on them if one had no idea what they were.

4.2.c Responsibility, Reflection, and Responsiveness to Reasons

According to the supervisory model, self-knowledge is essential for maintaining rational coherence in one's mental life. Furthermore, given that it is in virtue of our capacity to exercise reflective control over our mental states that we can be held responsible for them, self-knowledge is also essential to rational agency. The proponent of reflective control argues not only that our second-order beliefs about the reasonableness of our first-order states may serve as reasons for those states, but that, as far as our status as rational agents is concerned, they are the primary reasons that "rationally motivate" those

states.[52] This is not to say that, on this view, rational belief formation must always involve second-order reflection on the soundness of the reasons for it. A subject's belief that *p* – say, that a mouse has taken up residence in her house – may be based on a first-order awareness of pieces of evidence – a hole chewed in a bag of rice, what appear to be mouse droppings on the shelf – that serve as reasons that motivate and provide sufficient justification for the belief that *p*. However, if she is to be held responsible for her first-order state, the subject must be capable of forming a judgement concerning whether or not it is justified through second-order reflection on the justificatory force of the first-order belief and reasoning that supports it. Her focus is not on whether or not the hole in the bag and droppings were caused by a mouse, but rather whether or not her evidentiary beliefs about these things warrant a belief in its presence. And this higher-order judgement must determine whether or not she holds the belief that a mouse is indeed in the house.

What does this involve? Say a subject believes that *p* for reasons *q* and *r*. First of all, if she is to reflect on her belief that *p* and her reasons for it, she must know what that belief and reasons are – she must form true second-order beliefs about them. She then deliberates on the soundness of the first-order belief by examining those beliefs that serve as reasons for it, as well as the reasoning that connects them to it. This includes judging whether they themselves are justified, whether any fallacies in reasoning have been committed, and whether the evidence represented by those beliefs is sufficient to support

[52] The term 'rationally motivate' is also borrowed from Owens. It is meant to "register the fact that reasons for belief produce belief ... by explaining their product in a way that makes sense of it"(Owens, 17).

the belief they are taken to motivate.[53] Having successfully applied her knowledge of epistemic norms to her reasons and reasoning, she may either (1) find that everything meets the epistemic mark, upon which case she endorses the belief as one she ought to have and maintains it, or (2) find some fault in her reasoning and judge that she ought not hold the belief, at which point she changes her mind. In this way the subject assumes responsibility for her belief.

Even if taken at face value, this picture faces difficult questions. The proponent of reflective control claims that a subject's second-order judgement that her first-order *prima facie* reasons and reasoning in support her first-order belief that *p* are in order is what ultimately motivates her belief that *p*. So on this view, what directly motivates a subject's first-order belief that *p* for which she may be held responsible is not her first-order judgements about the world, but rather her second-order belief that the normative constraints on belief have been met. In other words, in light of her second-order judgement that the belief that *p* is sound and ought to be believed she decides to believe it. But, it may be asked, can such second-order judgement really play the motivational role envisioned for it? Owens points out that in order to reflect on the reasonability of her belief that *p*, the subject must already have a first-order awareness of the reasons that prompt that belief (Owens, 18). In exercising reflective control over her mental states, she engages in second-order judgement the purpose of which is to ensure her reasonability by explicitly acknowledging through that second-order judgement the normative force of the reasons she already has. But what do the subject's higher-order judgements that she has

[53] There is a threat of an infinite regress of justification here; however, I shall ignore this issue, as it is not the matter that generates the biggest problem for the view under consideration.

those reasons, and that they suffice for the reasonableness of her belief, add to the motivational equation? How do they exert an independent rational influence on – count as reasons for – her belief? As Owens puts it, "if you already have a non-reflective awareness of the reasons which ought to motivate you, how does the judgement that you ought to be moved by them help to ensure that you are so moved? Such judgements", he concludes, "look like an idle wheel in our motivational economy...." (ibid., 18). Indeed, this would seem to be the conclusion not only because such judgements merely confirm what is already the case. The mechanism of reflective control is second-order judgement about first-order states. The picture is of a mind turned inward, focused entirely on the rational standing of its own contents. How is it that the product of this inner inquiry – a mental state that refers to the epistemic fitness of other mental states – can serve as the primary reason to hold a first-order belief about the world?

The above questions about the motivational efficacy of second-order judgement about the epistemic standing of one's first-order belief presuppose that such judgements are possible – it is assumed that, through reflection on the reasons that rationally motivate her belief that *p,* the subject may arrive at a second-order belief that those reasons are (or are not) sufficient for that belief. And this second-order belief is what ultimately motivates the belief that *p* for which she may be held responsible. However, if a reasonable belief that *p* is one that is motivated by an awareness of sufficient evidence for that belief, then a problem arises. For, as Owens points out, reflection on strictly evidentiary beliefs that justify the belief that *p* may not determine whether or not those evidentiary beliefs are sufficient to rationally motivate the belief that *p* (Owens, 25). We may agree that the

formation of a rational belief that *p* or not-*p* should be determined by the balance of evidence for or against *p*. However, what determines what constitutes a sufficient level of evidence cannot be decided by deliberation on evidence alone. Rather, the point at which one judges that evidence to be sufficient for the formation of a belief will be determined partly by the subject's non-reflective sense of non-evidential considerations – for example, of the importance to the subject of the matter in question, or how much of his cognitive resources he is willing to devote to it. The fly in the ointment for the proponent of reflective control is that reflection on such justifying reasons (that is, that play a role in determining the rationality of a belief that *p*) cannot rationally motivate that belief. One cannot rationally motivate oneself to believe that *p* by reflecting on one's beliefs that time is running out and that it is important that a decision on whether *p* gets made.

The possibility of reflective control assumed by the supervisory model of self-knowledge depends upon the idea that if *r* is a reason to believe that *p*, and awareness of *r* rationally motivates the belief that p, then the second-order belief that one has reason *r* should also serve as a (indeed, *the*) reason that rationally motivates the belief that *p*. The objections just outlined suggest this cannot be so. I suggest that the moral of the story is that the formation of rational beliefs (those responsive to and governed by reason) must be a wholly first-order affair. The combination of the subject's first-order awareness of inconclusive evidence and the sort of non-evidentiary pragmatic considerations mentioned above work together to provide her with what is, from her point of view, sufficient reason to rationally motivate her belief that *p*.

4.2.d First-Order Reasoning and the Rational Adjustment of Mental States

The denial that second-order deliberation and the self-knowledge it presupposes may figure in the rational motivation of our mental states is consistent with the expressivist understanding of the character of self-ascriptions outlined earlier. On this view, we need not engage any higher-order cognitive faculty to reliably self-ascribe our mental states; rather, this capacity is explained by first-order linguistic expressive know-how. The suggestion is that just as our ability to self-ascribe our mental states is what one might call a first-order accomplishment, so too is our ability to maintain a rationally coherent mental life.

Suppose one believes that p, but that one is confronted with evidence that constitutes a *prima facie* reason for a contradictory belief that q. According to the supervisory model of self-knowledge and rationality, the adjustment of beliefs in light of this new evidence would require a host of second-order beliefs, among them beliefs that one has these competing beliefs, beliefs about their inconsistency, and beliefs about what changes in those beliefs would be required to resolve the discrepancy between them (Shoemaker: 1996c, 33). Once again, instead of being directly responsive to the first-order reasons to believe that q that the evidence presents (by either being convinced of the truth of q and thereby relinquishing the contradictory belief that p, or taking that evidence as reason to engage in further first-order inquiry concerning whether p or q), the subject must elevate his focus onto his own psychological states, as opposed to the state of the world, and through consideration of them alone arrive at a judgement about what first-order belief he ought to hold. Now, as mentioned in the earlier discussion of Burge, one problem here is

that this presupposes that the subject's higher-order reasons to which he is responsive are themselves sound, and so the proponent of this view owes an explanation of how this might be ensured that does not trigger an infinite regress of additional layers of overseeing judgement. A second problem is that, even assuming they were true, whatever second-order beliefs I arrive at regarding the content of my first-order beliefs and how they relate to one another would be irrelevant to the determination of whether or not the particular first-order beliefs in questions were true. And if what justifies me in holding a belief must be something relevant to whether the belief is true, then those second-order beliefs ought (also) to be irrelevant to whether I am justified in holding that first-order belief and – to that extent – to its rationality.

Let's look briefly again at Burge's argument for our entitlement to self-knowledge. He starts with the assumption of the fact of second-order critical reason and the role it plays in the overall maintenance of our rationality. From this he argues that the rational efficacy of this second-order deliberation – that it fulfils its contributing role – depends upon its also being in accord with reason. This, he argues, constitutes the source of our entitlement to those second-order beliefs. For, as mentioned earlier,

> ... if one lacked entitlement to judgements about one's attitudes, there could be no norms of reason governing how one ought to check, weigh, overturn, confirm reasons or reasoning. For if one lacked entitlement to judgements about one's attitudes, one could not be subject to rational norms governing how one ought to alter those attitudes given that one had reflected on them. (Burge: 1998c, 249)

The idea is that rationally necessary adjustment of one's mental states grounded on reasons and reasoning requires second-order deliberation on those mental states. However, as Moran emphasises, what should determine whether or not a rational subject forms a belief or desire that *p* should be the first-order reasons for or against the truth or desirability of *p* (Moran, 93). In his rational deliberation on what he ought to believe the subject must be responsive to first-order reasons. And this, the suggestion is, is just the mode the subject is in when engaged in supposedly higher-order deliberation on what he ought to believe. In weighing, overturning, or confirming his reasons or reasoning the subject engages in the same sort of first-order deliberation that produced the mental states "under review," and is subject to the same rational norms to which we appeal when describing the rational standing of those first-order states. As Owens argues in a slightly different context, he is subject to or governed by reason in virtue of the fact that he is responsive to first-order reasons (Owens, 18).

A final comment: As alluded to in the earlier discussion of Burge (see 2.2j), even if the supervisory model of rationality and the view of self-knowledge it entails were sound, it would still retain a limited attractiveness as an explanation of first-person authority. For while it could account for the authoritative self-ascription of those states that are responsive to reasons, such as beliefs and desires, it would have nothing to say in explanation of the similar degree and kind of authority we have with respect to self-ascriptions of sensations of pain, emotions, and the like.

4.2.e A Bottom-Up View of Rationality – Davidson and Radical Interpretation

On the supervisory model, second-order deliberation and the self-knowledge it presupposes are essential to maintaining rationality. It presents a "top-down" view of rationality – the subject, equipped with knowledge of rational norms and her first-order states, must be able to apply those norms to those states to ensure rational order. I have raised questions about the plausibility of this view on four fronts. First, I have argued in favour of an expressivist reading of authoritative self-ascriptions that challenges the idea that we have the sort of self-knowledge that the supervisory model presupposes. Second, I have suggested that, following Owens, even if we had the kind of self-knowledge (in particular, knowledge of our beliefs concerning non-evidential considerations that figure in the rational motivation of belief), that knowledge could not be deployed in the formation of the second-order beliefs about what we ought to believe. Third, I have proposed that, as Moran points out, it is first-order reasons to which we are rationally beholden when forming first-order beliefs, intentions, and other first-order mental states. Finally, I have suggested that the denial of a role for self-knowledge in (supposedly) second-order deliberation about what one ought to believe is consistent with an expressivist reading of those authoritative self-ascriptions thought to be involved in it. In other words, it will come as no surprise to the expressivist – who denies that authoritative self-ascriptions express second-order beliefs – that self-knowledge cannot play the kind of role envisioned for it by the advocates of reflective control. For as the expressivist sees it, we lack that sort of self-knowledge in the first place.

I would like to conclude this discussion with a brief look at an alternative approach to the supervisory understanding of rationality, namely that of Donald Davidson. This discussion is meant as a first-step towards relieving potential anxieties about rationality that might arise in light of the loss of the supervisory model. For it might be thought that without second-order supervision, and the control over our mental lives that comes with it, we would fall into irrationality. But on Davidson's model, and any other model that can eschew the supervisory "top-down" approach, there is no basis for such anxiety – rationality comes in at the ground floor, along with our first-order mental states; our being rational is part and parcel of our being first-order believers.

According to Davidson, rationality, thought, and speech are interdependent phenomena, the understanding of which requires that we focus on the communicative situation and what is needed for a hearer to successfully interpret the words of a speaker. We have seen how this informs his understanding of first-person authority. To briefly recap, Davidson combines two ideas:

(1) the semantic externalist claim that the meaning of a person's words "depends in the most basic cases on the kinds of objects and events that have caused the person to hold the words to be applicable; similarly for what the person's thoughts are about" (Davidson: 2001b, 37),

with

(2) the regularity thesis, namely that "whatever objects and events a person regularly applies her words to – i.e., whatever way they are regularly used – gives them the

meaning they have (and her thoughts the content they have as expressed by her use of those words)" (ibid., 37-8).

Together, (1) and (2) explain how a speaker cannot misunderstand the words she uses to express her mental states. These two theses are essential to the possibility of interpretation – without this basic connection to the world, and the regularity of use that is required for the individuation of that connection, there would be nothing for the interpreter to interpret. In other words, these two facts do not merely provide essential clues that serve as a way into the meanings of a speaker's words – they are part and parcel of what it is to speak meaningfully.

The same sort of thinking informs his understanding of two other interrelated ideas key to his view of rationality, namely the holism of the mental and the Principle of Charity. According to the former, a single belief, desire, or intention that *p* depends for its identity on the relations it bears to a host of other propositional attitudes. As he summarises it with respect to beliefs,

> Because of the fact that beliefs are individuated and identified by their relations to other beliefs, one must have a large number of beliefs if one is to have any. Beliefs support one another, and give each other content. Beliefs also have logical relations to one another. As a result, unless one's beliefs are roughly consistent with each other, there is no identifying the contents of beliefs. A degree of rationality or consistency is therefore a condition for having beliefs. (Davidson: 2001g, 124)

Given that every other propositional attitude depends for its identity on a great many beliefs, the point extends to the whole range of mental states. As Davidson writes a little

further on, "[t]here are ... no beliefs without many related beliefs, no beliefs without desires, no desires without beliefs, no intentions without both beliefs and desires" (ibid., 126).

Contrast this with the supervisory model of rationality. On that view, a failure of rationality equals a failure to effectively monitor and control one's first-order states through second-order deliberation on them. Were this failure to occur, one would still have such first-order states, however irrational that would make one. But on Davidson's holistic model, that would not be possible; sufficient disarray would preclude the possibility of assigning anything like such states to the individual. But if so, then the supervisory model would seem to presuppose a problematically high degree of atomism for our mental states.

Given the interconnected nature of mental states, the interpreter makes her way into the speech and thoughts of another holistically, as opposed to atomistically – it is an ongoing (in fact, for Davidson, never-ending) process whereby light dawns gradually (and, over time, more fully) on the whole. From the beginning of this process the interpreter must deploy the Principle of Charity. This principle

> calls on us to fit our own propositions (or our own sentences) to the other person's words and attitudes in such a way as to render their speech and other behaviour intelligible. This necessarily requires us to see others as much like ourselves in point of overall coherence and correctness – that we see them as more or less rational creatures mentally inhabiting a world much like our own. (Davidson: 2004, 35)

The assumption that those we seek to understand are rational by our standards must be in play from the outset – without it, the interpretive process could not get off the ground. Thus, it would be a mistake to see this principle, as some have, as a merely heuristic policy intended to counsel the interpreter on how to choose between competing possible (in the sense of minimally plausible or reasonable) interpretations.[54]

This is where the Principle of Charity first comes in. The interpreter must work to match up sentences of the speaker's language with the observable conditions in which they are uttered. She then devises a claim, expressed in her own language, about the truth conditions of that utterance ('"Le chat est sur le tapis" is true if and only if the cat is on the mat'). But to do so, the interpreter must take it that, in general, the speaker is getting her world right – utters true sentences that cohere with one another (true and coherent by the interpreter's lights, as Davidson often adds). From the interpreter's point of view, without this constraint there would be no reason to suppose that the behaviour observed constituted meaningful speech, that there was anything there to interpret. Put another way, there would be no grounds upon which the interpreter could differentiate a radically mistaken ascription of belief from a more plausible (reasonable) one given the same evidence (utterance plus behaviour in a given surroundings). Insofar as "anything would go" in this regard, interpretation could not get off the ground. Thus, the assumption of charity is a condition for the possibility of interpretation, and thus of a creature's counting as having speech and thought. As Davidson says, "[t]he policy of rational accommodation or charity in interpretation is not a policy in the sense of being one

[54] See Ramberg, pp. 71-77, for a detailed discussion of this matter.

among many possible successful policies. It is the only policy available if we want to understand other people." Thus, he continues, "[w]e should not think of this as some sort of lucky accident, but as something built into the concepts of belief, desire, and meaning" (Davidson: 2004, 36).

Looked at this way, Davidson offers what in comparison to the supervisory model might be called a "bottom-up" conception of rationality. According to the supervisory model, a fully rational subject is one who can regulate her mental life through second-order deliberation on the rational standing of ontologically distinct first-order states – rationality is a function of the subject's capacity to deploy knowledge of rational norms in the analysis those states, through which she is able to exercise rational self-control. This picture suggests a conceptual compartmentalisation of the mental that the holism entailed by radical interpretation eschews. For Davidson, intentionality arises out of the communicative situation understood in terms of radical interpretation. On this approach, the analysis of the concept of rationality reveals its intrinsic relation to the interconnected concepts of belief, desire and meaning. From the standpoint of radical interpretation, finding a speaker rational is a necessary condition for coming to an understanding of her utterances and knowledge of her mental life. In a sense, the communicative situation imposes rationality on us – the possibility of communication, and of finding one another as having mental lives at all, depends upon each communicator's interpreting her interlocutor's utterances in such a way that they conform to her norms of rationality which, if they succeed in communicating, they must share.

As mentioned above, this brief discussion can only be considered a potential starting point for the discussion of how to conceive of rationality without self-knowledge. Even so, it should assure us that worries over the possibility of rationality that might arise from the loss of substantive self-knowledge and the kind of higher-order control presupposed by the supervisory model of self-knowledge are premature.

4.3 Conclusion

I opened Chapter 1 with a review of the features of our self-ascriptions (their immediacy, groundlessness, and security) that have been at the center of many recent discussions of self-knowledge. I noted a recent trend amongst a number of philosophers' attempts to explain these distinctive features by pointing out what they take to be necessary ties between our capacity for self-knowledge and rational agency. In Chapter 2 I took a closer look at the views of three philosophers (Shoemaker, Burge, and Bilgrami) who subscribe to a particular understanding of this general connection, viz. what I have called the supervisory model of rationality. Problems with each account suggested the need for an alternative approach. Thus, in Chapter 3, I first examined Richard Moran's case for an epistemically "substantial" yet non-perceptual understanding of introspective self-knowledge that made its own distinctive appeal to the requirement of self-knowledge for rationality. I suggested that certain problems with his account of "genuine avowal" – that is, of self-ascriptions expressive of self-knowledge gained from the first-person perspective – actually point toward a non-epistemic account of such self-ascriptions.

Thus, I turned to the explanation of one such version – that of Donald Davidson – that I argued provides the basis for a full account of our capacity for authoritative self-ascription. However, I suggested that to complete the explanation, Davidson's account requires the kind of supplementation that an expressivist understanding of self-ascriptive utterances provides. This brought us to Chapter 4.

Three tasks were accomplished in this final chapter. First, there was the completion of the epistemically deflationary account of first-person authority begun in Chapter 3 with Davidson's insights regarding semantic authority. I showed how, contrary to Hacker's expressivist critique of Davidson, expressivism could actually be employed to fill out Davidson's explanation of first-person authority. This involved explaining two key ideas: First, that contrary to many philosophers' understanding of expressivism, the truth-evaluability of self-ascriptions need not confer assertoric status on them; consequently, a self-ascription could serve to both report and express a self-ascribed mental state. This paved the way for understanding the second key point, namely the basis of the connection between sincerity and truth in our self-ascriptions: If a speaker has the first-order state she expresses by way of a present-tense self-ascription – that is, if she is sincere – then that self-ascription will be true. I then defended this account against various objections, one of which (that of Dorit Bar-On) is motivated in part by the perceived need to make room for the sort of connection between self-knowledge and rationality that the supervisory model of rationality presupposes.

Thus, with a deflationary account of first-person authority in place, I then turned to the second matter of how such a non-epistemic account might affect that supervisory

model. What happens to it if we lack the kind of self-knowledge on which it depends? In a sense, nothing; for, following Owens, I argued that the supposed higher-order judgements about the rational standing of our first-order states by which we are said to exercise control over our beliefs and related states (and which presuppose self-knowledge of those states) cannot do the job assigned to them. Since second-order beliefs cannot serve as reasons to hold a first-order state, they are, as Owens puts it, an idle wheel when it comes to the rational motivation of belief. But neither should we expect them to play such a role. For, in line with part of Moran's discussion of self-knowledge, I argued that our deliberations about what we ought to believe, desire, and intend should be guided by our understanding of the first-order reasons for them. Finally, I noted that all of this was consistent with a deflationary account of first-person authority that denies that we have the sort of self-knowledge thought to be necessary by the advocate of the supervisory model of rationality.

This brought me to the third task of the chapter: addressing what a non-supervisory account of rationality – or rationality without self-knowledge – might look like. To this end I offered a synopsis of Davidson's argument that, in light of the holism of the mental, we ought to think of rationality as intrinsic to the concepts of belief, desire, and meaning. But if true, then there is neither room nor need for the supervisory model of rationality.

Bibliography

Bar-On, D. (2004). Speaking My Mind. Oxford: Oxford University Press.

Bar-On, D. and Long, D. (2001). "Avowals and First-Person Privilege." Philosophy and Phenomenological Research. Vol. 62, No. 2, March: pp. 311-355.

Bilgrami, A. (1999) "Self-Knowledge and Resentment." In C. Wright, B. Smith, and C. MacDonald (eds.) Knowing Our Own Minds (pp. 207-241). Oxford: Oxford University Press.

Blackburn, S. (1990). "Wittgenstein's Anti-Realism." In J. Brandt and R. Haller (eds.), Wittgenstein: Towards a Re-evaluation (pp. 13-26). Vienna: Holder-Pickler-Tempsky.

Boghossian, P. (1998) "Content and Self-Knowledge." In P. Ludlow and N. Martin (eds.), Externalism and Self-Knowledge (pp. 149-173). Stanford: CSLI Publications. Originally published in Philosophical Topics, 17 (1989), pp. 5-26.

Brueckner, A. (1998). "Shoemaker on Second-Order Belief." Philosophy and Phenomenological Research. Vol. LVIII, No. 2, June: 361-364.

_____. (1999). "Two Recent Approaches to Self-Knowledge." *Nous*. Vol. 33, Supplement: Philosophical Perspectives, 13, Epistemology: pp. 251-271.

Burge, T. (1998a). "Individualism and the Mental." In P. Ludlow and N. Martin (eds.), Externalism and Self-Knowledge (pp. 21-83). Stanford: CSLI Publications. Originally published in P.A. French, T.E. Uehling, and H.K. Wettstein (eds.), Midwest Studies in Philosophy: Studies in Metaphysics, Vol. 4, (1979), pp. 73-121.

_____. (1998b). "Individualism and Self-Knowledge." In P. Ludlow and N. Martin (eds.), Externalism and Self-Knowledge (pp. 111-127). Stanford: CSLI Publications. Originally published in The Journal of Philosophy, Vol. 85, No. 11, (1988), pp. 649-663.

_____. (1998c). "Our Entitlement to Self-knowledge." In P. Ludlow and N Martin (eds.) Externalism and Self-Knowledge (pp. 239-263). Stanford: CSLI Publications. Originally published in Proceedings of the Aristotelian Society, New Series, Vol. 96 (1996), pp. 91-116.

_____. (1998d). "Reason and the First Person." In C. MacDonald, B. Smith, and C. Wright (eds.), Knowing Our Own Minds (pp. 243-270). Oxford: Oxford University Press.

_____. (1982). "Two Thought Experiments Reviewed." Notre Dame Journal of Formal Logic, Vol. 23, Number 3, pp. 284-293.

Cassam, Q. (ed.). (1994). Self-Knowledge. Oxford: Oxford University Press.

Davidson, D. (1984). "On Saying That." In Donald Davidson, Inquiries into Truth and Interpretation. Oxford: Oxford University Press. Originally Published in Synthese, 19 (1968-9), pp. 130-146.

_____. (2001a). "A Coherence Theory of Truth and Knowledge." In Donald Davidson, Subjective, Intersubjective, Objective. Oxford: Oxford University Press. Originally published in D. Henrich (ed.), Kant oder Hegel?, (Stuttgart: Klett-Cotta, 1983).

_____. (2001b). "Knowing One's Own Mind." In Donald Davidson, Subjective, Intersubjective, Objective. Oxford: Oxford University Press. Originally published in Proceedings and Addresses of the American Philosophical Association, Vol. 60, No. 3 (Jan., 1987), pp. 441-458.

_____. (2001c). "First-Person Authority." In Donald Davidson, Subjective, Intersubjective, Objective. Oxford: Oxford University Press. Originally published in Dialectica, Vol. 38 (1988), pp. 101-111

_____. (2001d). "The Myth of the Subjective." In Donald Davidson, Subjective, Intersubjective, Objective. Oxford: Oxford University Press. Originally published in

_____. (2001e). "What is Present to the Mind." In Donald Davidson, Subjective, Intersubjective, Objective. Oxford: Oxford University Press. Originally published in The Mind of Donald Davidson, Grazer Philosophische Studien, Vol. 36 (Amsterdam: Rodopi, 1989).

_____. (2001f). "The Second Person." In Subjective, Intersubjective, Objective. Oxford: Oxford University Press. Originally published in P. French, T.E. Uehling and H. Wettstein (eds.), Midwest Studies in Philosophy, 17 (Indianapolis: University of Notre Dame Press, 1992).

_____. (2001g). "The Emergence of Thought." In Donald Davidson, Subjective, Intersubjective, Objective. Oxford: Oxford University Press. Originally published in Erkenntnis. Vol. 51, No. 1, (Sept., 1999), pp. 511-521.

_____. (2001h). "Afterthoughts." In Donald Davidson, Subjective, Intersubjective, Objective. Oxford: Oxford University Press.

_____. (2003). "Responses to Barry Stroud, John McDowell, and Tyler Burge." Philosophy and Phenomenological Research. Vol. 67, No. 3, Nov., pp. 691-699.

_____. (2004). "Expressing Evaluations." In Problems of Rationality. Oxford: Oxford University Press. Originally published as a Lindley Lecture, University of Kansas (1984).

_____. (2005). "The Social Aspect of Language." In Truth, Language, and History. Oxford: Oxford University Press. Originally published in B. McGuinness and G. Oliveri (eds.), The Philosophy of Michael Dummett, (Dordrecht: Kluwer Academic Publishers, 1994).

Evans, G. (1982). The Varieties of Reference . Oxford: Oxford University Press.

Finkelstein, D. (2000). "Wittgenstein On Rules and Platonism." In Alice Crary and Rupert Read (eds.), The New Wittgenstein (pp. 53-73). London: Routledge.

Fricker, E. (1998). "Self-Knowledge: Special Access versus Artefact of Grammar — A Dichotomy Rejected." In C. Wright, B. Smith, and C. Macdonald (eds.), Knowing Our Own Minds (pp. 155-206). Oxford: Oxford University Press.

Hacker, P.M.S. (1986). Insight and Illusion: Themes in the Philosophy of Wittgenstein. Oxford: Clarendon Press.

_____. (1997). "Davidson On First-Person Authority." The Philosophical Quarterly 47, no. 188, July: 285-304

Hamilton, A. (2000). "The Authority of Avowals and the Concept of Belief", European Journal of Philosophy 8:1 (2000), pp. 20-39.

Heal, J. (2003). Mind, Reason, and Imagination: Selected Essays in Philosophy of Mind and Language. Cambridge: Cambridge University Press.

Jacobsen, R. (1996). "Wittgenstein On Self-Knowledge and Self-Expression." The Philosophical Quarterly. Vol. 46, No. 182, January: 12-30.

_____. (2007). "Truth and Sincerity: Self-Knowledge Without Epistemology", Queens University, February 2007.

Moran, R. (2001). Authority and Estrangement: An Essay on Self-Knowledge. Princeton: Princeton University Press.

Owens, D. (2000). Rationality Without Freedom. London: Routledge.

Putnam, H. (1975). "The Meaning of 'Meaning'." In Mind, Language and Reality. Cambridge: Cambridge University Press. Originally published in K. Gunderson (ed.), Language, Mind, and Knowledge, Minnesota Studies in the Philosophy of Science, VII (Minneapolis: University of Minnesota Press, 1975).

Ramberg, B. (1989). Donald Davidson's Philosophy of Language: An Introduction Oxford: Basil Blackwell.

Shoemaker, S. (1996a). "On Knowing One's Own Mind." In Sydney Shoemaker, The First-Person Perspective and Other Essays. Cambridge: Cambridge University Press. Originally published in J. E. Tomberlin (ed.), Philosophical Perspectives, 2, Epistemology, (Atascadero: Riverview Publishing Co., 1988), pp. 183-209.

_____. (1996b). "First-Person Access." In Sydney Shoemaker, The First-Person Perspective and Other Essays. Cambridge: Cambridge University Press. Originally published in J. E. Tomberlin (ed.), Philosophical Perspectives, 4, Action Theory and Philosophy of Mind, (Atascadero: Riverview Publishing Co., 1990), pp. 187-214.

_____. (1996c). "Self-Knowledge and 'Inner Sense'." In Sydney Shoemaker, The First-Person Perspective and Other Essays. Cambridge: Cambridge University Press. Originally published in Philosophy and Phenomenological Research, LIV (1994), pp. 249-314

_____. (1996d). "Moore's Paradox and Self-Knowledge." In Sydney Shoemaker, The First-Person Perspective and Other Essays. Cambridge: Cambridge University Press. Originally published in Philosophical Studies, 77 (1995), pp. 187-214.

Siewart, C. (2003). "Self-Knowledge and Rationality: Shoemaker on Self-Blindness." In B. Gertler (ed.), Privileged Access: Philosophical Accounts of Self-Knowledge (pp. 131-145). Burlington: Ashgate Publishing Ltd.

Sinnott-Armstrong, W. (1994). "The Truth of Performatives." International Journal of Philosophical Studies, Vol. 2: pp. 99-107.

Smith, M. (1994). "Why Expressivists About Value Should Love Minimalism about Truth." Analysis. Vol. 54, No. 1, January: pp. 1-11.

Strawson, P.F. (1959). Individuals. London: Methuen.

Taylor, C. (1985). "The Concept of a Person." In Charles Taylor, Human Agency and Language: Philosophical papers, Vol.1. Cambridge: Cambridge University Press.

Wittgenstein, L. (1963). Philosophical Investigations. Edited by G.E.M. Anscombe and R. Rhees. Translated by G.E.M. Anscombe. Oxford: Blackwell Publishers.

Wright, C. (1992). Truth and Objectivity. Cambridge, Mass.: Harvard University Press.

_____. (2001a). "On Making Up One's Mind." In Crispin Wright, Rails to Infinity. Cambridge, Mass.: Harvard University Press. Originally published in P. Weingartner and G. Schurz (eds.), Logic, Philosophy of Science and Epistemology, the Proceedings of the XIth International Wittgensteinian Symposium. (Vienna: Holder—Pickler-Tempsy, 1987), pp. 391-404.

_____. (2001b). "Wittgenstein's Rule-Following Considerations and the Central Project of Theoretical Linguistics." In Crispin Wright, Rails to Infinity. Cambridge, Mass.: Harvard University Press. Originally published in A. George (ed.), Reflections on Chomsky. (Oxford: Blackwell, 1989).

_____. (2001c). "Wittgenstein's Later Philosophy of Mind: Sensation, Privacy and Intention." In Crispin Wright, Rails to Infinity. Cambridge, Mass.: Harvard University Press. Originally published in K. Puhl (ed.), Meaning Scepticism. (New York: de Gruyter, 1991).

_____. (2001d). The Problem of Self-Knowledge (I)." In Crispin Wright, Rails to Infinity. Cambridge, Mass.: Harvard University Press.

_____. (2001e). The Problem of Self-Knowledge (II)." In Crispin Wright, Rails to Infinity. Cambridge, Mass.: Harvard University Press.