Philosophy and
Phenomenological Research

Philosophy and Phenomenological Research
Vol. LXXVIII No. 1, January 2009

Two Kinds of Self-Knowledge

MATTHEW BOYLE
Harvard University

I argue that a variety of influential accounts of self-knowledge are flawed by the assumption that all immediate, authoritative knowledge of our own present mental states is of one basic kind. I claim, on the contrary, that a satisfactory account of self-knowledge must recognize at least two fundamentally different kinds of self-knowledge: an active kind through which we know our own judgments, and a passive kind through which we know our sensations. I show that the former kind of self-knowledge is in an important sense fundamental, since it is intimately connected with the very capacity for rational reflection, and since it must be present in any creature that understands the first-person pronoun. Moreover, I suggest that these thoughts about self-knowledge have a Kantian provenance.

1. Introduction

Kant famously held that we possess two fundamentally different kinds of self-knowledge: knowledge of ourselves through "inner sense" and knowledge of ourselves through "pure apperception."[1] The former faculty, he claimed, gives us knowledge of our own sensations, knowledge of ourselves as passive beings, while the latter gives us knowledge of what we think and judge, knowledge of our own "spontaneity" (cf. B67-8, A107, B132, B153, B278). Nevertheless, although he held that these two faculties are distinct, he also thought there was a relation of dependency between them: he argued that knowledge of ourselves through inner sense would be impossible in the absence of a capacity for pure apperception

[1] For the distinction, see *Critique of Pure Reason*, A107, B132 and B152-9. Kant sometimes calls inner sense "empirical apperception," distinguishing it from the "pure apperception" expressed by the representation "I think" (see A107, B132). At other times, however, he simply uses the term "apperception" to name the nonempirical form of self-awareness (see, e.g., B153). I will follow the latter practice: when I speak of "apperception," I will mean *pure* apperception, the kind distinct from inner sense. All subsequent references to the first *Critique* are by the pagination of the first ("A") and second ("B") editions. I quote from the Kemp Smith translation (Kant 1929). When not referring to the first *Critique*, I cite the volume and page number of the standard *Akademie* edition of Kant's works (1900, cited as "Ak.").

(at, e.g., B140). For, he claimed, the latter capacity, the one manifested in our ability to think of ourselves as "I," is the capacity that makes us "knowing beings" at all (cf. A117n, B132, B157n).

These Kantian claims are, of course, obscure, and the interpretative literature that has grown up around them is vast.[2] Without worrying about how to interpret Kant's actual views, however, we can think of the claims just mentioned as constituting the schema for a possible view about self-knowledge. Such a view would distinguish two kinds of knowledge of our own minds, an active and a passive. But although it would hold that these kinds of self-knowledge are distinguishable, it would also maintain that there is a relation of dependency between them—that one of the two kinds could not exist without the other.

I mention this schema for a view about self-knowledge not with the aim of doing Kant exegesis, but because I think seeing the possibility of such a view can help us to resolve a dispute in the contemporary literature on self-knowledge.[3] There have recently been a number of important attempts to account for our ability to know our own minds by connecting this ability with our capacity for some kind of agency. The most developed of these accounts is Richard Moran's recent *Authority and Estrangement* (2001), which argues that our ability to know our own current beliefs, desires, and other attitudes can on at least some occasions be understood as reflecting an ability to "make up our minds:" an ability to know our minds by actively shaping their contents.[4] This sort

---

2 Recent books on the topic include Ameriks 2000, Brook 1994, Keller 1998, Kitcher 1990, and Powell 1990.

3 Following a widespread practice, I will use the term "self-knowledge" to refer to the awareness expressed in a subject's ability to speak in the first person, without self-observation and with apparent authority, about her own present mental states. Arguably, we also have this sort of knowledge of certain facts about, e.g., the current position of our own limbs and about what we are doing. Exactly what sorts of facts can be known in this special way, and what sort of privilege or "authority" belongs to such knowledge, are questions that will be discussed in the body of the paper. For the moment, I simply use "self-knowledge" as a label for whatever sort of knowledge it is that has interested philosophers under such headings as "privileged access," "first-person authority," and so on. I take it for granted that our ability to speak authoritatively about our own present mental states does reflect *knowledge*. This has been denied, but I do not know of any convincing argument for denying it, and surely it deserves to be the default position. I say something in defense of this presumption in §4.

4 Other writers on self-knowledge who give a central role to the notion of agency include Burge 1996 and 1998 and Bilgrami 1998. There has also been considerable recent interest in another kind of connection between agency and self-awareness: interest, namely, in how our capacity to think of ourselves in the first person is connected with our capacity, not to make up our minds, but actually to intervene bodily in the world at large. This is not my topic here, although I will touch on it at the end of §5. For recent discussion of these issues, see the essays collected in Eilan and Roessler 2003.

of view has come under criticism, however, for its inability to account for our immediate, authoritative knowledge of other kinds of mental states that do not seem to be subject to our active control. I want to argue that this criticism rests on a false assumption about the uniformity of self-knowledge, an assumption that overlooks the possibility of a view like Kant's. And indeed, once we have this possibility in mind, I think its attractions will be obvious.

I begin, in §2, by sketching Moran's account and how an appeal to the notion of agency figures in it. Moran admits that there are kinds of authoritative self-knowledge to which his account does not apply, but he claims that the kind of self-knowledge it does describe is fundamental. Given his admission about the limited scope of his account, however, it is hard to see what grounds he can have for this claim of fundamentality. Indeed, critics of Moran have taken the fact that there are species of self-knowledge to which his account does not apply as itself reason for thinking that we must look for some other account. In §3, I show that this criticism depends on the assumption that a satisfactory account of self-knowledge should be fundamentally uniform, explaining all such knowledge in the same basic way. This assumption is accepted uncritically by many writers on self-knowledge, writers whose views are otherwise quite dissimilar.

The remainder of the paper criticizes this assumption and lays the groundwork for a different kind of view. In §4, I argue that a satisfactory account of self-knowledge must account for the distinction between behavior that merely manifests the presence of a certain mental state and behavior that expresses a representation of oneself as in a state of the relevant sort. I then show, in §5, that only an account which recognizes a distinction between the kind of self-knowledge Moran characterizes and the kind we have of what we are presently sensing can account for this distinction. The upshot is that we must recognize two kinds of self-knowledge, one exemplified in our knowledge of our own sensations, the other in the kind of knowledge Moran investigates—our knowledge of our own judgments. I conclude, in §6, with some remarks about the Kantian character of this outlook.

2. Moran on Self-Knowledge and Agency

There are familiar reasons to be puzzled by our capacity for self-knowledge. It is widely recognized that we can know our own present thoughts, attitudes, and sensations in a way that is fundamentally different from the way we know of the mental states of other persons. The precise character of this difference is a matter of dispute, but it is generally agreed that (1) a person is normally in a position knowledgeably to

ascribe various kinds of mental states to himself without needing the sorts of *evidence* that would be required for his ascription of such states to another person, and that (2) self-ascriptions of these kinds of mental states are not normally liable to the same kinds of *error* that afflict ascriptions of such states to other people. The former feature of the relevant ascriptions is commonly referred to as their *immediacy*, while the latter is one manifestation of their *authority*—their apparent entitlement to some sort of deference not accorded to third-person, evidence-based ascriptions of the same kinds of states. The *general problem of self-knowledge* is to explain how we can be in a position to speak about our own minds in such an immediate and authoritative manner, while still counting as speaking about the very same states that can be known to others only on the basis of observation or inference.

Moran's orientation, however, is not primarily toward this general problem but toward a particular subdivision of it. Early in his book, he remarks that

> [t]here are two basic categories of psychological state to which the ordinary assumption of 'privileged access' is meant to apply: occurrent states such as sensations and passing thoughts, and various standing attitudes of the person, such as beliefs, emotional attitudes, and intentions. (I will have comparatively little to say here about the case of sensations, which I believe raises issues for self-knowledge quite different from the case of attitudes of various kinds.) (2001, pp. 9–10)

Moran's view thus seems to be that the general problem of self-knowledge really comprises two different problems, one having to do with "attitudes" and the other with "sensations." Moreover, his decision to focus on our knowledge of our own attitudes, while leaving aside our knowledge of our own sensations, suggests that he regards these two problems as to a significant extent independent of one another.

What Moran finds striking about our knowledge of our own attitudes is this: we often seem to be able to know whether we hold them *by deliberating about the topics they concern*. If I want to know whether I believe that p, it seems that I can normally answer this question by considering whether there is *reason* to believe that p—whether there are persuasive grounds for thinking that p is true. And analogously, at least for so-called "motivated desires," it seems that I can normally determine whether I want X by considering whether there is reason to want X—whether there are persuasive grounds for thinking that X would be desirable. Likewise for intending to do something, for hoping for something, for fearing something, and so on: although there are certainly cases in which such attitudes prove recalcitrant in the face of our reflection on reasons for and against, still it is striking that often

enough we can simply say what our attitude is by deliberating about the topic in question. Moran has popularized the term "transparency" as a label for this kind of relationship between a question about whether I hold a certain attitude and a question about the object of that attitude.[5] Thus the question whether I believe that *p* is said to be "transparent" to the question whether *p* because it seems that I can settle the former question by settling the latter.

Such transparency can seem strange: how can I justifiably answer a question about what I believe by answering what seems to be a quite different question about a state of affairs independent of my belief? Moran's master thought is that the way to explain this seemingly paradoxical situation is to understand such self-knowledge as involving a kind of agency: I can know whether I believe that *p* by deliberating about whether *p* because my deliberation about *p* can constitute my *making up my mind* to believe that *p*. Thus, in a recent article summarizing his position, he explains his view as follows:

> What right have I to think that my reflection on the reasons in favor of P (which is one subject-matter) has anything to do with the question of what my actual *belief* about P is (which is quite a different subject-matter)? Without a reply to this challenge, I don't have any right to answer the question that asks what my belief [about, e.g., whether it will rain] is by reflection on the reasons in favor of an answer concerning the state of the weather. And then my thought at this point is: I *would* have a right to assume that my reflection on the reasons in favor of rain provided me with an answer to the question of what my belief about the rain is, if I could assume that *what* my belief here is was something determined by the conclusion of my reflection on those reasons. (Moran 2003, p. 405)

Moran's thought, in short, is that our ability to speak authoritatively about our own beliefs without looking for signs of belief in our behavior becomes intelligible if we suppose that to conclude that *p* on the basis of deliberation normally just amounts to coming to believe that *p*, and that a subject who possesses the concept of belief will understand that this is so.[6] If a subject who possessed the concept of belief were entitled to assume that, in reaching the conclusion that *p* is true,

---

[5] See Moran 2001, Chapter 2, §6. Moran cites Edgley 1969 as the source of the term.

[6] This is not to suggest that my merely taking myself to have a certain belief must make it so. As Moran emphasizes, his view does not demand that a subject be incorrigible about her own attitudes, or that her claims about her own attitudes have special authority no matter what their basis. What is important is that the question of what attitude I hold *can* often enough be settled by me on the basis of deliberation about whether *p*, and that—according to Moran—it is fundamental to the very possibility of thought about such attitudes that this should be so.

he was coming to believe that p, then it seems that he could justifiably answer the question whether he believed that p by reflecting on grounds for taking p to be true. And it seems appropriate to describe such knowledge of one's own beliefs as reflecting a kind of agency, for the subject's concluding that a certain proposition is true would be what made it the case that he believed the relevant proposition.

If we grant that Moran has described a possible form of self-knowledge, it is natural to ask how much of our actual self-knowledge takes this form.[7] Moran himself invites this question by suggesting, at various points in his book, that his observations shed light not only on our knowledge of our own beliefs but on our capacity for privileged self-knowledge in general. At one point, for instance, he characterizes himself as having

> argued the case for seeing the ability to avow one's belief as the fundamental form of self-knowledge, one that gives proper place to the immediacy of first-person awareness and the authority with which its claims are delivered. (2001, p. 150)

And in another passage he claims that the capacity to make "transparent" attitude-ascriptions is "what makes the difference between genuine first-person awareness and a purely theoretical or attributional knowledge of one's own states" (2001, p. 107). It is difficult, though, to see how these sorts of claims can be defended. There seem plainly to be kinds of mental states of which our knowledge is both

---

[7] Another sort of question Moran's account must face concerns how the exercise of deliberative agency secures *knowledge* for the deliberating subject. If Moran's story is to explain, not just how a subject can come to hold a certain belief, but how he can come to know that he holds it, then it must give some account of how his conviction that he believes that p is warranted; and it seems that Moran's appeal to "making up one's mind" is meant to supply the relevant account. But how does my making it the case that I believe that p warrant me in believing that I believe that p? At least on some readings of "make it the case" and some substitutions for p, it seems that I might make it the case that p without knowing that p: for instance, I might make it the case that my hair is on fire by bending too close as I light the stove, but nevertheless (unfortunately) not come to know that my hair is on fire. If, by contrast, "making up one's mind" does normally supply one with knowledge of what one believes, how it can do so needs to be explained. (For this sort of concern about Moran's account, see for instance O'Brien 2003 and Wilson 2004.)

I think this challenge raises issues of great interest, but I do not attempt to address them in this paper. The issue that concerns me here is why the sort of power exercised in "making up one's mind" should be regarded as fundamental, and what implications its fundamentality has for the shape an account of self-knowledge in general must take. When I speak of what deliberation provides us with as "self-knowledge," I do so on the assumption that some satisfactory account of the epistemology of agential knowledge can be given, but the points I am going to make do not depend on this. I hope to address issues about the epistemology of deliberative agency in future work.

non-observational *and* non-deliberative: not just sensations but, for instance, appetites (i.e., brute, unreasoned desires for things of a certain kind) and what might be called "recalcitrant attitudes" (e.g., feelings of anger that I know to be unjustified but cannot overcome). Moran does not deny that we can have authoritative first-person knowledge such attitudes.[8] But if his account does not apply to such knowledge, it is not clear how he can justify his claim to have described the fundamental form of self-knowledge, the one that "makes the difference" between first-person awareness and the kind of awareness we might have of the mental states of another person.

Indeed, given that we can often speak authoritatively about our own attitudes without going through any process of conscious deliberation, it is not clear how much light Moran's account sheds even on our knowledge of the kinds of mental states that are his principal topic. If I am asked such questions as whether I believe that Washington crossed the Delaware or whether I want to come along to the beach, I can often just answer straightaway, without any reflection on grounds for and against. In such cases, although I am surely expressing immediate and authoritative knowledge, the knowledge does not seem to reflect my now exercising the power to make up my mind. My mind, it is natural to say, is already made up. The question what I believe or desire is still, of course, transparent for me to a question about what is so or what is desirable, but the relevant convictions of fact or desirability are not being formed in the present, and so it is hard to see how an appeal to agency can help to explain my present knowledge of them. This sort of observation has led some philosophers to distinguish between the transparency of the question whether *to* believe that *p* to the question whether *p* and the transparency of the question whether I *already* believe that *p* to the question whether *p*. Thus, in a recent paper, Nishi Shah and J. David Velleman argue that the explanation of the relevant transparency must be quite different in the two cases: in the former case, it is a matter of my being able to make up my mind by thinking about whether *p*; in the latter, a matter of my putting the question whether *p* to myself "as a stimulus applied to [my]self for the empirical purpose of eliciting a response."[9] If this distinction is sound, then it seems that the application of Moran's agency-based account of self-knowledge is in fact quite limited.

It is thus difficult to see how Moran could justify his claim to have described the fundamental form of self-knowledge. Even if we grant that

---

[8] See for instance his remarks on "unmotivated desires" which are not "an expression of one's reasons" at Moran 2001, p. 115, his remarks on unconquerable jealousy and fear at pp. 58 and 63, and his discussion of recalcitrant belief at pp. 131–132.

[9] See Shah and Velleman (2005), p. 506.

the kind of self-knowledge he describes is *an* immediate and authoritative kind of knowledge, and grant also his claim that a subject must at least *sometimes* take herself to be capable of knowing at least *certain* of her attitudes in this way,[10] still it is not clear why this shows that this sort of self-knowledge is more fundamental than other forms, or what its immediacy and authority has to do with the immediacy and authority with which we know various other kinds of mental states. This much, however, is clear: if there is truth in Moran's claim to have described the fundamental form of self-knowledge, it must be because "the fundamental form" does *not* mean "the form an account of which can serve as the model for an account of all immediate, authoritative self-knowledge." My own view is that there is a sense of "fundamental" on which Moran's claim is true. In §5, I will try to take some steps toward clarifying this sense. First, though, I need to take issue with a widespread assumption which holds the untenable reading of "fundamental" in place.

3. An Assumption Underlying Criticisms of Moran

In a number of recent responses to Moran's book, the observation that his account applies only to certain kinds of mental states, and even to those states only on certain occasions, is urged as a reason for rejecting the account, or at least for rejecting Moran's claim to have described the fundamental form of self-knowledge. In her recent *Speaking My Mind* (2004), for instance, Dorit Bar-On argues that, although Moran's view may account for the kind of knowledge of our own attitudes that we achieve through deliberation about the topics of those attitudes, it

---

[10] This is argued in the provocative but difficult fourth chapter of Moran's book. At the conclusion of this chapter, Moran remarks that

> [t]he problem with the idea of generalizing the theoretical stance toward mental phenomena is that a person cannot treat his mental goings-on as just so much data or evidence about his state of mind all the way down, and still be credited with a mental life (including beliefs, judgments, etc.) to treat as data in the first place. (2001, p. 150)

At least part of the argument for this seems to be that, even when I take a "theoretical stance" toward some aspect of my mental life—i.e., even when in a given case I look for behavioral evidence that I hold a certain attitude—still in doing so I necessarily presume that I can in general make up my mind on the basis of evidence. Hence, although I can treat the existence of a given attitude as a "mere datum," I cannot in general doubt my power to make up my mind on the basis of grounds without implying, absurdly, that I am incapable of even entertaining the question what my attitudes are (compare Moran 2001, p. 148). This seems right, but it is difficult to see how to get from here to the conclusion that a creature which did not treat its attitudes as open to deliberation could not be "credited with a mental life." I shall argue that the proper conclusion to draw is that such a creature could not be credited with self-knowledge.

cannot account for "what is distinctive about avowals traditionally so-called—i.e., ordinary, present-tense self-ascriptions of occurrent states of mind," since "this feature is only characteristic of a certain subclass of avowals" (p. 131). Consequently, she suggests, Moran's account fails to meet one of the basic desiderata for an account of self-knowledge, namely that "it should apply to intentional and non-intentional avowals alike, and allow us to separate the various types of avowals from other ascriptions" (p. 144). Similarly, in his *Expression and the Inner* (2003), David Finkelstein observes that "we speak with first-person authority about a great many mental states and events that are not avowable in Moran's sense" (p. 162), and concludes that Moran's claim to have described the fundamental form of self-knowledge is unwarranted, since his account does not provide "a model … for explaining how we manage to speak with authority about a wide range of our own inner states and goings on" (p. 155).

The assumption underlying these criticisms is evidently that we should seek some common explanation of all of the cases in which we can speak immediately and authoritatively about our own mental states. We could call this *the Uniformity Assumption*, for it amounts to the demand that a satisfactory account of our self-knowledge should be fundamentally uniform, explaining all cases of "first-person authority" in the same basic way. Bar-On and Finkelstein are not alone in making this assumption. It is also made, for instance, in Shaun Nichols and Stephen Stich's recent *Mindreading* (2003), which defends an account of self-knowledge quite different from the ones advocated by Bar-On and Finkelstein—an account that appeals to hypothesized "monitoring mechanisms" that supply us with appropriate second-order beliefs about our own first-order mental states. Although they do not address Moran's view specifically, Nichols and Stich do criticize what they call "ascent routine strategies" of accounting for self-knowledge—accounts that explain our ability to say whether we believe that *p* in terms of our mastering an "ascent routine" which tells us that we can move from an answer to the question whether *p* to and answer to the question whether we believe that *p*. The appeal to such ascent routines is, they argue, "clearly inadequate as a general theory of self-awareness" (2003, p. 194), since there are many kinds of self-knowledge which could not be arrived at by such an ascent routine. But like Bar-On and Finkelstein, Nichols and Stich simply assume that a theory of self-awareness *should* be general—that it should give some single, undifferentiating account of all knowledge which exhibits some sort of first-person privilege.

Nor are these authors atypical. The Uniformity Assumption is arguably present wherever philosophers are content to speak of *the* way in

which we know our own minds. Anyone familiar with the literature on "first-person authority" will recognize that this sort of outlook is widespread.[11] To be sure, the assumption is rarely stated explicitly, but it is evident in the common tendency to argue about whether our immediate, authoritative knowledge of our own mental states is to be accounted for by appeal to some sort of quasi-perceptual faculty of "inner sense,"[12] by a reliable tendency of our second-order beliefs about our own mental states to track our first-order mental states,[13] by a linguistic convention that simply grants a person's psychological self-ascriptions some sort of default authority,[14] or by the fact that such ascriptions normally "express" the states that they report.[15] What defenders of all these views have in common, despite their differences, is the conviction that *some* single basic strategy of explanation will account for all cases of immediate, authoritative self-knowledge. This is the conviction I want to question.

4. A Minimal Condition on Self-Knowledge

To see why the Uniformity Assumption should be rejected, it will be helpful first to reflect on a minimal requirement that a subject must meet if any of her utterances are to count as expressing knowledge of her own mental states: namely, the obvious requirement that she must *understand* whatever sentences she uses to express this knowledge. Although this condition is obvious, it is worth emphasizing

---

11 Widespread but certainly not universal. The suggestion that we possess different kinds of self-knowledge which are to be accounted for differently appears, for instance, in Davidson 1984, Shoemaker 1990, Burge 1996, Bilgrami 1998, and Falvey 2000. But although these authors propose accounts of self-knowledge that apply only to a restricted class of mental states, and thus presuppose the falsity of the Uniformity Assumption, they tend not to give principled reasons why an account of self-knowledge *must* treat different sorts of states differently. Perhaps for this reason, these interventions have not altered the shape of the mainstream debate.

12 This sort of view is more often discussed than defended. It is often attributed to some giant of modern philosophy such as Descartes, Locke, or Kant. My own view is that, at least in Kant's case, although he does speak of a faculty of "inner sense," it is a travesty to attribute to him the view that the knowledge of our own mental states supplied by this faculty is quasi-perceptual. To defend this interpretative claim, however, would require another paper.

13 Defenses of this sort of view include Armstrong 1968, Lewis 1972, and Lycan 1998.

14 The most influential version of this approach is presented in a series of papers by Crispin Wright: see his 1987, 1991, and 1998.

15 Early examples of this approach can be found in Ryle 1949 (e.g., at p. 102) and Shoemaker 1963, Chapter 6. (Shoemaker has subsequently adopted a more complex position, which shares Moran's emphasis on transparency. See his 1988 and 1990.) More recent defenses of the expressivist approach include Jacobsen 1996, Hamilton 1998, Bar-On and Long 2001, Finkelstein 2003 and Bar-On 2004.

because it will turn out that a subject can only meet it if she meets certain further conditions which are less obvious, and which are frequently overlooked in discussions of self-knowledge. In particular, I shall argue that this requirement on the expression of self-knowledge reflects a more basic requirement on self-knowledge *per se*: a self-knower must *represent* her own condition as being of a certain kind. And it will emerge in the next section that only an account of self-knowledge that recognizes the distinctness and fundamentality of the kind of self-knowledge identified by Moran can account for the relevant sort of representation.

To see how the capacity to express self-knowledge depends on a capacity for self-representation, it is helpful to contrast the kind of ability possessed by a speaker who can use a self-ascriptive sentence to express knowledge of her own mind with the kind of ability that might be exhibited by a suitably-trained parrot. Suppose I train a parrot to cry out "I'm in pain!" just when it is, in fact, in pain. Then it will be disposed to utter a self-ascriptive sentence on just those occasions when the sentence is true; and surely it is at least as plausible to say in the parrot's case as in the case of a human speaker that the utterance is not made on the basis of inference or observation. But we also want to say something else, namely that *the parrot does not understand what it is saying*: it utters a form of words with a certain conventional content, but it does not grasp this content. And this implies that the parrot's vocalizations cannot express knowledge of its own pain in the way that similar sentences might in the mouth of a competent speaker. For the parrot's utterances do not express a classification of its condition as one which these words aptly characterize. Rather, insofar as it has merely been conditioned to cry out "I'm in pain!" when it is in pain, what it has is merely a learned addition to whatever repertoire of behaviors parrots naturally have for expressing pain, and the new behavior expresses pain only in the sense in which the various other behaviors do. But surely the natural pain-expressing behaviors of a parrot just manifest pain itself, not the parrot's *knowledge that* it is in pain. We should reserve the latter sort of ascription for cases in which a creature acts in a way that manifests, not just pain, but grasp *that* it is in pain—i.e., grasp that a certain subject is in a certain kind of condition. For whatever else it requires, knowing that p presumably requires representing that p, and only where a creature's activity expresses the attribution of a certain property to a certain subject is there a ground for saying that it has any *representation* of its condition at all.

Now, a competent speaker who sincerely avows "I'm in pain" plainly does express a representation of his own condition. We should

admit this even if we deny, as many writers on self-knowledge do, that one normally has some independently-specifiable ground for claiming that one is in pain. Even if normal avowals of pain are in this sense groundless, they are surely not blind in the way that a parrot's vocalizations are blind: a subject who sincerely and comprehendingly says "I'm in pain" must understand what "pain" is and take himself to be in such a state. Perhaps children learning to avow pain go through a phase in which their use of self-ascriptive sentences is like that of our imagined parrot, but when a mature competent speaker sincerely avows "I'm in pain," he does so *because he takes this to be the case*. To suppose otherwise would be, not to explain how we can have authoritative knowledge of our own mental states, but simply to deny that we do have such knowledge: it would amount to reclassifying the statements that apparently express self-knowledge as mere automatic responses, which perhaps entitle an observer, or the subject himself, to judge that he is in a certain mental state, but which do not themselves express such a judgment.

If we are to take seriously the idea that our avowals can express *knowledge* of our own minds, then, we must distinguish between two senses in which a kind of behavior might be said to "express" a mental state: the sense exemplified in the utterances of our imagined parrot, which we might call *the manifestation sense* ($\text{expression}_{\text{M}}$), and the sense exemplified in the superficially similar utterances of a competent speaker, which we might call *the representation sense* ($\text{expression}_{\text{R}}$). Talk of the expression of *self-knowledge* is in place only where there is behavior that $\text{expresses}_{\text{R}}$ the relevant mental states, since otherwise the behavior in question is sufficiently accounted for by the mental states themselves, without appeal to the subject's knowledge of them. I have used an example of linguistic behavior to bring out this point, but the point can be accepted even by someone who supposes that creatures which do not speak a language can manifest knowledge of their own minds. What is crucial is not that the creature should express its self-knowledge in an articulate language but that *whatever* sort of activity is supposed to manifest this knowledge should have a certain kind of explanation: one that adverts, not merely to the creature's *being in* the mental state supposedly known, but to the creature's *representing* its own state as of a certain kind.

I do not think that many theorists of self-knowledge would dispute these points: the capacity for self-knowledge is generally assumed to involve the capacity for self-representation. But although this is conceded by nearly everyone, I think the point has consequences which are not well appreciated. For most writers proceed as if this can be conceded while leaving entirely open what shape an account of

self-knowledge should take. We can see how this might be a problem by briefly considering a kind of account of self-knowledge favored by several of Moran's critics: namely, the sort of "expressivist" account which holds, roughly, that we are capable of avowing our own present mental states without relying on observation or inference because our normal avowals are not reports but "expressions," which stand to the mental states they express in relations analogous to the relation in which crying stands to pain. The ability to cry when one is in pain, and to do so "without observation or inference," is not mysterious, for crying is not a report on one's condition but a manifestation of an unreflective disposition which one has when one is in pain. Similarly, contemporary expressivists suggest, the ability to make verbal avowals of one's own present mental states without observation or inference would be unmysterious if such avowals were expressions rather than reports. Of course, expressivists admit, there is a great difference between crying and saying "I'm in pain," for the latter but not the former is a *linguistic* expression of pain, capable of being true or false. But still, they suggest, the linguistic expression might be a learned addition to our repertoire of pain-expressing behaviors, standing in a relation to pain analogous to that of natural expressive behaviors.[16]

The relevance of the foregoing discussion to this sort of view should be clear. I have argued that if a self-ascription of pain is to express self-knowledge, it must not merely $\text{express}_{\text{M}}$ but $\text{express}_{\text{R}}$ pain. It seems clear, however, that the natural disposition to cry when in pain merely $\text{expresses}_{\text{M}}$ pain; so a subject who merely learned to utter "I'm in pain" where he had formerly been disposed to cry would so far be exhibiting no more self-knowledge than is exhibited by our imagined parrot. When expressivists say that avowals of present mental states are "linguistic" expressions "capable of being true or false," they must mean that they are not just parroted tokens of phonetic types that count as self-ascriptive sentences in a certain language. They must mean that the relevant tokenings are actually comprehending $\text{expressions}_{\text{R}}$ of the subject's condition. But on the face of it, explaining how the relevant utterances can be $\text{expressions}_{\text{R}}$ introduces complications that make it difficult to preserve any simple analogy with the relation between crying and pain: this relation seemed unmysterious precisely because it involved the operation of an unreflective behavioral disposition; but if an utterance is to $\text{express}_{\text{R}}$ a creature's mental state it must

[16] This is a very brief, but I hope not an inaccurate, summary of a line of thought to be found in authors such as Jacobsen 1996, Hamilton 1998, Bar-On and Long 2001, Finkelstein 2003 and Bar-On 2004.

not be an operation of this kind, for then it would only express$_M$ the relevant state.[17]

In the next section I will argue that explaining how a creature's utterances can express$_R$ its mental states requires that we credit it with a faculty which supplies it with a special kind of knowledge of its own deliberated attitudes, a kind of knowledge that has the structure Moran describes. This kind of knowledge, however, is different in principle from the kind we have of our own sensations. If this is right, then a satisfactory account of self-knowledge must recognize more diversity than expressivists typically do. But the point does not bear only on expressivists. Any theory of self-knowledge will confront an analogue of the challenge I have posed for expressivism: it must leave room for an account of what it is not just to *have* mental states but to *represent* one's own mental states. If I am right that accounting for this requires crediting the subject with a special kind of knowledge of his own deliberated attitudes, then any theory of self-knowledge must leave room for knowledge of this special kind. But expressivists are not the only ones who attempt to give a general account of self-knowledge without giving detailed attention to what is involved in representing one's own mental states. This is also typical, e.g., of philosophers who suggest that our capacity to know our own mental states reflects "monitoring mechanisms" which reliably produce higher-order representations of our first-order mental states. Defending such a view, William Lycan remarks that

> if [such a] theory is false, that is a brutely empirical fact; certainly Mother Nature could have equipped us with banks of first- and second-order internal monitors, whether or not She did in fact choose to do so. (1998, p. 758)

But just as we can ask what it is for an utterance not merely to express$_M$ but express$_R$ a mental state, so we can ask what it is for one mental state to "monitor" another mental state not merely in the sense in which the level of mercury in a thermometer monitors temperature (monitoring as manifesting) but in the sense of representing the relevant state *to the subject* (monitoring as representing). Lycan's suggestion that what accounts for self-knowledge in general might be Mother

[17] Of course contemporary expressivists are not silent on the question of what is involved in the transition from "natural behavioral expressions" of mental states to linguistic expressions of such states. But in no defense that I am aware of is there an explicit recognition that the notion of expression itself must be understood in a different sense where the relevant expression is linguistic, or an explicit consideration of how the capacities drawn on in linguistic expression might themselves be connected with the power to know one's own mind.

Nature's having equipped us with "internal monitors" reflects his assumption that an account of what it is for one mental state to be a higher-order representation of another mental state will not set constraints on how we must account for self-knowledge. I shall now argue that this assumption is mistaken.

5. Judgment, Reasons, and Self-Knowledge

In §2, we observed that Moran's claim to have described the fundamental form of self-knowledge is defensible only if "fundamental" means something other than "capable of serving as the model for an account of self-knowledge in general." In §3, we saw that many approaches to the problem of self-knowledge in effect assume that this is what "fundamental" must mean: they assume that an account of self-knowledge must apply uniformly across the board. As we have just seen, however, such approaches tend to take for granted the idea of a subject's representing her own mental states, without detailed investigation of the preconditions of such representation. I now want to argue that the power to represent one's own mental states presupposes the power to know one's own deliberated attitudes in the way that Moran specifies. If this is right, then the kind of self-knowledge Moran describes must find a fundamental place in any satisfactory account of self-knowledge. And if, as everyone agrees, Moran's account cannot serve as a model of self-knowledge in general, then it follows that we must reject the Uniformity Assumption.

My argument for these conclusions comes in three stages. I first discuss what it is for a subject to be able to represent her own mental states, and how this power is connected with the ability to speak about one's own mental states (§5.1). Next I argue that a subject with the relevant sort of representational power will necessarily be entitled to draw conclusions about what she believes in the way that Moran describes (§5.2). This part of the argument turns only on general considerations about what it is to be the kind of subject who has the kind of representational powers that would explain comprehending speech, and it establishes only a point about such a subject's *entitlement* to draw conclusions about her beliefs in a certain way: it does not show that she must understand this entitlement. In the final subsection, however, I argue that a person who can represent her own mental states must understand that the subject to whom she ascribes those states is one who has the power to make up her mind (§5.3). This part of the argument turns on a point, not about the presuppositions of comprehending representation in general, but about the presuppositions of *self*-representation in particular.

5.1

Before proceeding, I need to clarify what I mean when I speak of a subject's representing her own mental states. There are, after all, various things one might call "self-representation." In one sense, a psychological theory might posit "subpersonal" self-representations which figure in the explanation of how the brain is able to perform some computational task such as calculating the location of objects in its vicinity on the basis of information about stimulations of its retina. Perhaps explaining such an ability requires positing states which have the function of representing, e.g., the stage of processing that has been reached, and this might be called a self-representation. The point of calling such representations "subpersonal," however, is to indicate that they belong, not to *the subject's* view of what is so, but to a level of representing of which the subject might be quite oblivious. The kind of representation that is a condition of authoritative self-knowledge must be a self-representation in a stronger sense: it must be a representation of the subject's mental state *which is predicable of the subject herself* rather than merely of some hypothesized processing mechanism operating within her. But even this is not enough, for, famously, I can represent what is in fact my own state but fail to represent it *as* my own state. Thus, in John Perry's well-known example, I might represent that *somebody* is pushing a shopping cart containing a torn bag of sugar but fail to represent that *I* am pushing that shopping cart, even though I am in fact the person in question.[18] The kind of self-representation that is a condition of authoritative self-knowledge must be a self-representation in a yet stronger sense: it must be a personal-level representation of the subject's mental state *as her own* mental state.

We can express these points in another way by saying that the kind of self-representation that is of interest to us is the kind that a subject with the relevant linguistic abilities (1) would be able to report in (2) an utterance involving a form of the first person. The former point captures the requirement that the representation form part of the subject's view of what is so, the latter that the subject should recognize that she herself is the object of the representation. We can characterize self-representation in this way without assuming that only a language-user can be a self-knower. Whether or not these abilities can be possessed by a languageless creature, it is clear that *one* way in which they can manifest themselves, in a creature that does speak a language, is in

[18] See Perry 1979. It requires a more contrived example to make the point in the case of mental states, but we can imagine, e.g., my seeing what is in fact my own reflection and wondering what *that person* is thinking while not wondering what *I* am thinking.

articulate speech. Moreover, if we are interested in understanding the sort of self-knowledge that mature language-users characteristically possess—namely, the ability to *say*, without observation or inference, what they presently think, see, feel, want, and so on—we should insist that our account of self-knowledge suffice to explain this ability. No doubt there are definable kinds of self-knowledge which a creature might possess without being able to articulate that knowledge in speech, but an account that caters only for such self-knowledge will not yet explain the kind of ability that is the traditional focus of discussions of "first-person authority," "privileged access," and so on—namely, the ability to *say* (avow, declare, etc.) what my own present mental state is. So at least for the purposes of understanding *this* sort of self-knowledge, we can focus on the question: what kinds of abilities must be exhibited in the comprehending linguistic expression of a self-representation?

5.2

Let us, then, return to our earlier comparison between a subject who comprehendingly says "I'm in pain" and a parrot whose utterance of "I'm in pain" expresses pain itself, but not a representation of its condition as of a certain kind. What sorts of abilities must we suppose the comprehending speaker to have that the parrot lacks? This much seems clear: a subject's utterance of "I'm in pain" expresses a representation of his condition, in the relevant sense, only if his producing it reflects an understanding of the meaningful elements out of which the sentence is composed. Whereas our imagined parrot merely produces a self-ascriptive sentence as an undifferentiated block, a comprehending speaker must produce the same sentence in such a way that his use of "I" expresses a comprehending representation of himself, and his use of the phrase "am in pain" comprehendingly predicates a certain kind of mental state of that subject. Only if his speech-behavior reflects this sort of understanding are we entitled to say that he is not just reacting in a certain characteristic way to pain, but classifying his condition *as* of a certain kind.

In reflecting on the presuppositions of such understanding, we really face two questions. First: What is involved in understanding something as a complex representation composed from meaningful elements? Secondly: What is involved in understanding the particular elements that figure in a *self*-representation? I will turn to the latter question in the next subsection, but for the moment I want to concentrate on the former. I shall argue that the kind of power that would equip a creature to understand the content of its own

utterances would also entitle it to draw conclusions about its own beliefs in the way Moran describes.

What is involved in understanding the content of one's own utterances is a large question, but we can see something about it by thinking about what a comprehending speaker must be able to do, beyond merely producing various sentences, parrotwise, in conditions in which they are in fact apt. It is clear, I think, that at least part of what is required is that the subject should be able to reflect on relationships between the content of any given sentence and the content of various other sentences. For to suppose that her utterance of "I'm in pain" reflects an understanding of the meaning of "I" and "am in pain" is to suppose that her use of this sentence has a certain sort of explanation: one that adverts to more general abilities to use and understand sentences involving "I," on the one hand, and sentences involving the predicable "to be in pain," on the other.[19] If she cannot produce and understand other sentences involving these expressions, then her use of them in this combination does not manifest such abilities. And it is not enough that she should merely produce various sentences involving these expressions on appropriate occasions, or react appropriately to them: she must be able to recognize relationships between the truth of any one and the truth of others. For to suppose that her knowledge of what it is for a subject to be in pain, or of what it is to ascribe a property to herself, is exercised in her understanding of these various sentences is to suppose that her understanding of them depends on her recognizing a common element in them, and recognizing this element as common must involve the capacity to recognizing relationships (e.g., of implication, exclusion, and inductive support) between their contents. To have the ability to recognize such relationships, and adjust what one claims in light of them, is a precondition of understanding what one is doing, when one utters a sentence, as taking a stand on what is true, a stand related to various other stands she might take. But a subject who does not have this sort of understanding does not understand the content of her utterances, for any given content is the content that it is only in virtue of standing in such relations.

A comprehending speaker, then, must be able to make claims in a way that reflects a grasp of the relation of the content of any given claim to the contents of a system of possible other claims. And this in turn requires that she should, in general, be able to reflect on her

[19] These abilities will of course be conditional on her understanding the general structure of the relevant sentences and the other expressions they contain. But what we can say is this: if she understands the expressions "I" and "to be in pain" then, given a sentence involving these expressions, and provided that these other conditions are met, she should be able to understand that sentence.

grounds for holding a given claim true: for a subject could hardly be credited with the ability to grasp relations among various systematically related contents if her endorsement of any given content were not potentially open to modification by her consideration of its relation to other contents she endorses. We could summarize this requirement by saying that a comprehending speaker must be able to entertain a certain sort of "Why?"-question about the claims she makes, a question that asks for grounds that show the claim in question to be true. Or again, we could say that she must, as it is sometimes put, be able to "play the game of giving and asking for reasons," where "reasons" here means considerations bearing on the truth of the claims she has made.[20] It is by being able to engage in such reflection that a subject manifests, not just a disposition to use various sentences on appropriate occasions, but a grasp of the systematic relations among their contents. But now the thing to notice is that a subject capable of holding propositions true in a way that reflects an appreciation of this sort of "Why?"-question will also be entitled to ascribe beliefs to herself in a way that conforms to Moran's Transparency Condition. For a subject who can say that *p* just when she takes there to be sufficient grounds for supposing *p* to be true is a subject whose speech already expresses her beliefs: when she (nondeceptively) says "*p*," she will be affirming something she takes to be true, and since to take something to be true just is to believe it, she will also be entitled to say "I believe that *p*."

If this is right, then we are in a position to say why the kind of self-knowledge that Moran characterizes is fundamental. It is fundamental because the ability to say what one believes in the way that Moran specifies is intimately connected with the kinds of representational abilities that must be possessed by a subject who can make comprehending assertions, and a subject who lacks these sorts of abilities cannot be a self-representer, in the sense we have specified, at all. For on the one hand, as we have just seen, if a subject *can* make comprehending

[20] I borrow the phrase from Brandom 1994. I do not, however, mean the phrase to imply, as it does for Brandom, that this must be a game one plays *with other subjects*: I take no position on whether the capacity to reflect on reasons requires the capacity to communicate one's reasons to others. Even setting aside this commitment, the doctrine that the ability to make comprehending claims requires the ability to reflect on reasons is hardly uncontroversial. I think many philosophers would accept *something* like the views sketched in the last two paragraphs, but there would certainly be controversy over the details, and there are some philosophers who would reject the whole spirit of the view. Obviously a full-scale defense of this outlook cannot be attempted here: my aim is just to give a sketch of the kinds of considerations that speak in favor of it, and then to show its implications for an account of self-knowledge.

assertions, then she will necessarily be entitled to accompany the claims she thus expresses with "I believe." And on the other hand, if a subject *cannot* make comprehending assertions, then none of her apparently self-ascriptive utterances will express comprehending representations of her own states. Her utterances will be, at best, like those of our pain-avowing parrot, which utters true self-ascriptive sentences but does not understand them. A subject capable of expressing comprehending representations of her own mental states will thus be one who is entitled to make assertions about her own *beliefs* in the way that Moran describes. In short, if we ascribe to a subject the sort of representational power that can explain comprehending speech, then we at the same time attribute to that subject the kind of power that would allow her—provided that she understood the content of the question "Do I believe that *p*?"—to know her belief as to whether *p* by making up her mind.[21]

5.3

What I have said so far is only that a subject who has the sort of power of representation that can explain comprehending speech must be one who is *entitled* to accompany her sincere assertions with "I believe," not that such a subject must actually grasp this entitlement and be master of some expression which would allow her to capitalize on it. But this already implies that an account of self-knowledge must leave room for the possibility of the sort of self-knowledge that Moran describes, since it implies that any subject who can make comprehending assertions is one who *could* acquire self-knowledge through deliberation simply by coming to understand the relevant entitlement. And in fact I think there is an even closer connection between the possibility of deliberative self-knowledge and the possibility of self-knowledge in general. This becomes clear if we reflect, not just on the conditions of

[21] I have cast this as a point about the kind of representational power that could explain comprehending speech, but I think it could equally be cast as a point about the capacity for what philosophers have traditionally called *conceptual* representation. A concept is supposed to be a kind of representation our grasp of which equips us to understand what is common to an unlimited manifold of possible representations involving it. (Compare Kant's remark that "every concept must be thought as a representation which is contained in an infinite number of different possible representations (as their common character)" [A25/B40]; and for recent discussion of concept-possession in a similar spirit, see Evans 1982, Ch. 4, §3.) If grasp of a concept requires this sort of understanding, and if such understanding requires the further sorts of abilities that I have outlined, then the points I have made about the connection between Moran's sort of self-knowledge and comprehending speech are really points about the connection between self-knowledge and conceptual representation, and do not depend on special features of the ability to speak in particular.

comprehending representation in general, but on the conditions of self-representation in particular. We shall see that the connection that *we* observed in the last subsection—the connection between the power to make comprehending claims and the power to make up one's mind—is one that a subject capable of self-representation must *herself* have understood.

We noted in §5.1 that a claim is a *self*-ascription, in the interesting sense, only if it involves a form of the first-person. But there is a connection between understanding the first person and recognizing one's power to make up one's mind in the way Moran describes. For consider what qualifies an expression as a form of the first person. An expression "*A*" is a form of the first-person only if a subject who understands it understands that, in saying "*A* is *F*," he is predicating the property of being *F* of *himself*, i.e., the very person who is claiming that this predicate applies to this subject.[22] A subject who did not understand this would not yet be using the relevant expression to make *self-conscious* predications, even if he succeeded in referring to the person who was in fact himself. Now, we observed in §5.2 that to be able to make claims requires being able to make up one's mind about the truth of a given claim by considering grounds for and against. It follows that a subject's use of "*A*" will express self-consciousness only if it bears the right sort of connection to this ability: he must understand that the person he calls "*A*" is the very person whose mind is, so to speak, his to make up. This need not involve the readiness to articulate some abstract proposition about the content of the term "*A*": indeed, the subject may not have at his command the sorts of psychological

---

[22] This is just a restatement, in a linguistic register, of the traditional thought that the referent of the expression "I" is *the thinker*. The point of the traditional thought is this: to understand that the subject of a certain predication is *oneself* is to understand that the subject of the relevant predication is *the very subject who is thinking that this predicate applies to this subject*. This point is nicely expressed by Gareth Evans:

> [T]he essence of 'I' is *self*-reference. This means that 'I'-thoughts are thoughts in which a subject of thought and action is thinking about him*self*—i.e. about a *subject* of thought and action … I do not merely have knowledge of myself, as I might have knowledge of a place: I have knowledge of myself *as* someone who has knowledge and makes judgments, including those judgments I make about myself. (1982, p. 207)

Note Evans' suggestion that understanding the first person involves understanding oneself as the subject of thought *and action*. I think this is exactly right: although I have only been discussing knowledge of one's own beliefs and its connection with the ability to answer a certain kind of "Why?"-question, I think that similar points could be made about knowledge of one's own actions and its connection with the ability to answer a certain (different) kind of "Why?"-question. I say more about this shortly.

predicates that would allow him to articulate such a proposition. But even if he does not have command of such predicates, he must have an at-least-implicit grasp of the connection between his use of "*A*" and his capacity for deliberation.[23]

This is particularly plain if we consider how one's use of the first person must be linked to one's ability to make up one's mind to *do* something. A subject who judges "That plank is about to hit *A* in the head," and who has the normal aversion to being hit in the head, but whose so judging does not dispose him to take evasive action, is a subject whose use of "*A*" plainly does not express self-consciousness. By contrast, a subject whose use of "*A*" *is* connected in this sort of way with his decisions about what to do displays an awareness of the fact that the things he decides to do are the intentional actions of the thing he calls "*A*"—and this is so even if he does not possess a special expression whose role is to mark this awareness. We can imagine him learning that, whenever he decides to perform some action ϕ, he is also entitled to affirm "*A* intends to ϕ;" but whether or not he has learned this, his use of "*A*" reflects an understanding that his setting to do ϕ is the setting to do ϕ of the thing he calls "*A*."

Something similar applies in the case of making up one's mind that something is the case: a subject's use of "*A*" expresses self-consciousness only if he displays an awareness that his reaching the conclusion that *p* is the reaching of this conclusion by the thing he calls "*A*." For imagine a subject who was able to say "*p*" in the way we have described, on the basis of reflection on grounds for and against, but who was then uncertain whether the thing he called "*A*" was prepared to say that *p*.[24] Knowing that he is prepared to say that *p*, we ask him whether *A* would say that *p*, and he is unsure and needs to look for

---

[23] Compare G. E. M. Anscombe's well-known example of the "*A*"-language, in which "*A*" is a term which each speaker uses to refer to himself, but whose use does not "include self-consciousness" (Anscombe 1975, p. 50).

[24] For purposes of illustration, I assume that the subject has command of some predicate like "has said" or "would say," which links a speaker with an actual or possible claim-making utterance; but I do not insist that this is essential to understanding the first person. It is hard to imagine how a subject who does not possess some such predicate could use a term in a way that manifests understanding that the referent of that term is the very subject who is making the claim—which, as we have seen, is required if the term is to express self-consciousness. But if this *is* possible, then it should also be possible to construct cases in which a subject uses the relevant term without manifesting such understanding, and these cases would be enough to vindicate my point: that a subject uses a term as a form of the first person only if her use of the term is connected in the right way with her power to decide whether *p*.

behavioral evidence. Or on an occasion when he says that *p*, we ask him who has said so and he is unable to identify *A* as the speaker, or is able to do so only in an alienated way, as if his determination that "*p*" is true were one thing and somebody's just then uttering that very sentence were a surprising coincidence. Whatever saying "*A* is *F*" might signify for such a subject, it seems plain that it could not amount to his self-consciously predicating *F*-ness of the maker of that very predication. He would thus not be using "*A*" as a true first person. By contrast, a subject who *does* understand that his affirming "*p*" is *A*'s affirming "*p*" displays an awareness that when he decides that a proposition is true, this is a decision of the thing he calls "*A*"—even if he does not possess a special expression whose role is to mark this knowledge. We can imagine him learning that, whenever he decides that "*p*" warrants affirming, he is also entitled to affirm "*A* believes that *p*;" but whether or not he has learned this, his use of "*A*" reflects an understanding that his determinations of what is true are the determinations of the thing he calls "*A*."

If this is right, it sets an important constraint on the project of accounting for self-knowledge. We have seen that an account of self-knowledge must be an account of the subject's representing her own condition, and that the relevant sort of representation must be one whose linguistic expression would involve a form of the first person. But the upshot of our recent discussion is that a subject only understands the content of this sort of judgment if he understands that the subject of whom he predicates pain is a subject concerning whom he can know certain kinds of facts, not by observing *that they are* so, but by determining them *to be* so. Lacking such understanding, a subject can of course make utterances involving the English word "I;" but until he understands the link between his use of this term and his power to decide what is the case, he does not understand its significance. It follows that the kind of self-knowledge Moran describes is fundamental, not just in the sense that any self-knower must be in a position to *acquire* such knowledge by learning that he is entitled to attach "I believe that" to claims he is prepared to endorse, but in the further sense that, whether or not he has learned to use such an expression, his implicit grasp that he has the power to make up his mind is a condition of his understanding the first person at all. If his use of the term "I" does not reflect an understanding that it refers to the person whose mind is his to make up, he does not understand the content of this term, and hence does not understand the content of any sentence of which it is a part. And once again, the same point that applies at the level of speech applies, *mutatis mutandis*, at the level of thought: if a subject does not possess a representation which is linked, in the sort of way just described, to his power

to make up his mind about what is the case, then he does not possess the power of *self*-representation, and hence cannot entertain self-ascriptive thoughts. In particular, he cannot think thoughts about his own mental states. Hence he cannot be a self-knower.

It follows that Moran's sort of story about how we know our own beliefs must form a fundamental and independent part of any account of self-knowledge, whatever explanation it goes on to give of our capacity to know our own sensations, appetites, and so on. For the sort of self-knowledge Moran describes is a kind that must be available to any self-knower. This is not to say that any self-knower must actually have mastered some expression, like "believes" in English, that he can use to self-ascribe beliefs by attaching it to the conclusion of his reflection on whether *p*. But whether or not he has mastered such a predicate, he will have the deliberative capacity whose acts such a predicate serves to mark, and his use of the first person will reflect an understanding that he has this capacity.

An account of self-knowledge which accepts the Uniformity Assumption must either rule out the kind of self-knowledge Moran describes, or else maintain that all of our self-knowledge is of this kind. Everyone agrees that the latter option is untenable. But we have seen that no account of self-knowledge can afford to deny the possibility of Moran-type self-knowledge, for to deny it would be to deny a precondition of thought of *oneself*, and would thus undermine the possibility of self-knowledge in general. What Moran has given us, then, is not a model that can be generalized to account for all varieties of self-*knowledge*, but an account of the way of knowing one's own mind that is a precondition of self-*consciousness*. This kind of self-knowledge is fundamental because it characterizes the framework into which any story about other varieties of self-knowledge must fit.[25]

---

[25] Moran describes himself as concerned with our knowledge of our own "attitudes" in general, but apart from a few remarks about intention, I have focused exclusively on our knowledge of our own beliefs. Do the sorts of points I have made bear on our knowledge of our own attitudes more generally? In thinking about this question, it is helpful to distinguish two categories of attitudes: *cognitive attitudes*, which involve the holding-true of some proposition, and *conative attitudes*, which involve regarding some action as to-be-performed or some object as to-be-attained. (Perhaps there are attitudes that belong to both categories: nothing I say will exclude this.)

For cognitive attitudes, Moran's kind of self-knowledge is relevant precisely because these attitudes involve belief, and are open to deliberation in virtue of the fact that the relevant beliefs are open to deliberation. Thus, if it is true that propositional anger that *p* must involve the belief that the relevant fact constitutes an insult to me, or that propositional fear that *p* must involve the belief that the relevant fact constitutes threat, then such attitudes will be knowable transparently to

6. Conclusion: Two Kinds of Self-Knowledge

I have been arguing that an account of our knowledge of our own deliberated attitudes must form a fundamental and independent part of any satisfactory account of self-knowledge. I take it to be evident that a satisfactory account will also have to tell some other story about our knowledge of our own sensations and appetites: this is one point on which Moran and his critics agree. If this is right, then we must reject the Uniformity Assumption and admit that our ability to speak authoritatively about our own minds draws on two different kinds of self-knowledge.[26] I want to conclude by emphasizing two ways in which the resulting outlook resembles the Kantian view of self-knowledge that I mentioned at the outset.

One similarity between Kant's view and the view advocated here is that both draw a sharp distinction between the way we know our own judgments about what is the case and what to do, on the one hand, and the way we know our own sensations and appetites, on the other.

---

the extent that our capacity to deliberate about the relevant beliefs renders our anger and fear themselves open to deliberation. These attitudes will also, of course, involve passions that have a measure of independence from belief, and this begins to explain why such attitudes can prove recalcitrant in the face of reflection on reasons. To this extent, there is room for an appeal to some analogue of the expressivist story about pain in explaining our knowledge of such attitudes: learning to say when I am angry or afraid will no doubt involve learning to use words for what was formerly expressed by other kinds of behavior. But for such a thing as propositional anger or fear to be possible at all, these primitive stems of anger and fear must intertwine with our capacity for articulate belief, so that the resulting emotions at least normally bear a relation to our application of the concepts *insult* and *threat*.

In the case of conative attitudes such as desire and intention, the story is different but analogous. Grounds for conative attitudes are not reasons for thinking something true but reasons for thinking something desirable, reasons why there is "something to be said" for a given course of action. My capacity to know my own conative attitudes through deliberation is, like my capacity to know my own beliefs through deliberation, connected with my capacity to answer a certain sort of "Why?"-question; but the relevant "Why?" is different: it is, I think, the "certain sense of the question 'Why?'" that G. E. M. Anscombe investigates in her *Intention* (1963). Nevertheless, like our capacity to answer the truth-oriented "Why?"-question, our capacity to answer this practical "Why?"-question is arguably basic to our very status as rational thinkers and agents. But I will not attempt to argue for this here. My purpose is just to sketch how the points I have made about our knowledge of our own beliefs might be extended to the larger set of attitudes on which Moran takes his account to bear, and to indicate how this account might be *fundamental* to the understanding of our knowledge of our own deliberated attitudes without being a *total* account of such knowledge.

[26] No doubt further investigation might lead us to distinguish further kinds: my aim is not to suggest that there are *only* two kinds of self-knowledge, but to argue that there must be *at least* two, that *this* distinction at least is fundamental to any account of self-knowledge. It is fundamental because what motivates it is not merely the observation that the best explanation of self-knowledge requires that we treat our knowledge of different sorts of states differently, but a principled argument from considerations about what self-knowledge *must* involve.

A way of putting the thesis of the last section is to say that our immediate, authoritative knowledge of our own judgments is a necessary byproduct of our ability to reason about what is the case and what to do. It seems clear, though, that our knowledge of our own sensations and appetites is not in this sense maker's knowledge. However we explain our privileged knowledge of our own sensations and appetites, we should acknowledge that these are states that come to pass with us, not states we arrive at through deliberation. And this sounds strikingly like what Kant says: that whereas our apperceptive knowledge of our own judgments is a knowledge of "what we do [*thun*]," our knowledge of our sensations and appetites through inner sense is a knowledge of what we "undergo [*leiden*]."[27]

This Kantian contrast between an active and a passive form of self-knowledge has been a source of puzzlement to commentators. Our discussion, however, has equipped us to see a point in the distinction. For on the one hand, we have seen that it is attractive to understand our knowledge of what we believe as reflecting our capacity for a kind of agency—the capacity to make up our minds on the basis of grounds for belief.[28] The point of the invocation of agency here is not that every actual self-ascription of belief reflects a present *exercise* of the latter capacity. The connection is rather at the level of the capacities themselves: only a creature *capable* of making utterances in a way that is responsive to the sense of the question "Why?" that asks for grounds for holding-true can express its beliefs in comprehending assertions, and once this capacity is in place, the capacity to make immediate, authoritative *self-ascriptions* of belief requires only a harnessing of the former capacity to the use of an expression like "I believe." By contrast, our knowledge of our own sensations plainly is not connected in *this* way with reasons and deliberation: we do not, e.g., feel pain because we determine that there are sufficient grounds for so feeling. Our sensations are not in this sense up to us: they are, as Kant says, states we undergo.

Having made these observations, we are in a position to respond to a criticism of Moran's account mentioned toward the end §2: that often, when we ask ourselves whether we believe that *p*, we find our minds already made up. This is certainly true, but if it is supposed to show a deep difference between the way we know beliefs that we arrive

---

[27] The characterizations I am quoting are from Kant's *Anthropology*, §24 (Ak. 7:161). But the idea that apperception gives us knowledge of ourselves *qua* active (or "spontaneous"), while inner sense gives us knowledge of ourselves *qua* passive (or "receptive"), is common to all of Kant's discussions of the topic.

[28] Moran himself suggests that this thought has a Kantian provenance (see Moran 2001, Chapter 4, §7), although he does not discuss Kant's views in any detail.

at through present deliberation and the way we know beliefs that we simply call to mind, it misses the point of describing our knowledge of our own beliefs as a kind of active knowledge. The point is that, given the connection for me between the question whether I believe that *p* and the question whether *p*, and given the connection between my capacity to say whether *p* and my capacity to answer the truth-oriented "Why?"-question, it follows that my capacity to answer questions about what I believe is necessarily tied to my *capacity* to deliberate. We could put it this way: when I speak about what I believe, I am speaking about *how my mind is made up*—even if I am not *making* it up at the moment, even if I *never* went through a process of conscious deliberation about the belief in question. For my ability to make comprehending claims—even ones that are not the outcome of present deliberation—depends on my ability to make utterances in a way that reflects an appreciation of the truth-oriented "Why?"-question. This appreciation is exercised on those occasions when I actually make up my mind by deliberating, but it is present in the background even where no actual deliberation has taken place, since the relevant utterances only count as comprehending claims insofar as they reflect the *power* to confront this question. The relevance of this power is plain in the case where deliberation has actually taken place, but it remains real in cases where no deliberation has occurred, in virtue of the truth of counterfactuals about how I could have answered if the question "Why?" had come up.[29]

These remarks also bear on Nishi Shah and David Velleman's claim that, whereas I can answer the question whether *to* believe that *p* by making up my mind, I must answer the question whether I *already* believe that *p* by putting the question whether *p* to myself "as a stimulus applied … for the empirical purpose of eliciting a response." This does not seem apt as a description of the phenomenology of calling to mind a settled belief: Shah and Velleman's remark makes it sound as though I test my belief as to whether *p* as I might call into a well to see if there is an echo. If this were the case—if I just found the assertoric "*p*!" coming back when the

---

[29] Compare Anscombe's remark on the significance of Aristotle's account of the practical reasoning:

> [I]f Aristotle's account were supposed to describe actual mental processes, it would in general be quite absurd. The interest of the account is that it describes an order that is there whenever actions are done with intentions. (1963, §42, p. 80)

My point about the connection between belief and deliberation, similarly, is that what we see displayed in explicit deliberation about what to believe is an order that must be there whenever beliefs are held for reasons.

interrogative "*p*?" was sent in—then it is hard to see how this reply could figure in my present reflection as anything but the testimony of an alien voice, whose rational significance for my thinking now was an open question. But surely it is crucial that things are not normally like this: normally, the fact that I believe something, and know myself to believe it, has immediate rational significance for my present reflection, whether or not I am now forming the relevant belief in response to a deliberation.[30] Whatever way I have of retaining beliefs must retain their rational significance for me; otherwise it isn't *beliefs* that are being retained. And the explanation of how beliefs can retain their rational significance is that an extant belief that *p* is the same *kind* of thing as a newly formed belief that *p*: it is, as we have seen, my answer to the question whether *p*, formed and maintained in a way that is normally responsive to my reflection on grounds. To say this is not to deny that there can be beliefs that prove recalcitrant to reflection, perhaps even beliefs that we only come to recognize in ourselves through self-observation and self-analysis. It is only to state the rule against whose background such exceptions are intelligible.

This is all I will say about the first point of contact with Kant: the contrast between an active and a passive kind of self-knowledge. I want now to turn to a second point of contact: the thought that there is a relation of dependency between these two kinds of self-knowledge. Kant tells us that it must be possible for the apperceptive "I think" to accompany all of my representations if my representations are to be thinkable at all (B132). This implies, and it is clear from other passages that Kant holds, that the representations I receive through "inner sense," in particular, would not be thinkable by me if I did not have the capacity for apperception.[31] Let me restate this, without argument, in a way that makes its significance clearer: the claim is that I would not be able to think about the kinds of states that are the objects of inner sense—sensations, appetites, and other kinds of mental "affection"—if I did not also have (at least potentially) the distinctively active sort of awareness I have of my own thoughts and judgments.

Now, put this way, Kant's claim is again strikingly similar to a claim for which I have argued. For I have argued that our ability to self-

[30] I am indebted for the language of "immediate rational significance," and for the general shape of the point made here, to Burge 1996 and 1998.

[31] See B140, and compare the much-discussed letter to Marcus Herz of May 26, 1789 in which Kant claims that if I were a creature lacking the capacity for apperception, my representations would take place "without my knowing the slightest thing thereby, not even what my own condition was" (Ak. 11:52).

ascribe sensations depends on our ability to acquire immediate, authoritative knowledge of our own beliefs, since our very capacity to make comprehending assertions at all depends on our capacity to make utterances in a way that reflects an appreciation of the truth-oriented "Why?"-question, and our understanding of the first person, which must figure in any genuine *self*-ascription, depends on our understanding its connection with this capacity. My argument for the latter dependency turned on the thought that a subject has mastered the use of the first person only if he understands that it refers to the very subject who is making the claim, or thinking the thought, in which it occurs. And this again is a Kantian thought. Kant puts it in his way by saying that "the *I* is only the consciousness of my thinking" (B413). What he means is that this element of thought and intelligent speech has its significance in virtue of its connection with a certain kind of knowledge, the knowledge we have of our own thoughts in thinking them, in making up our minds.

A central point of this paper has been that, unless we recognize this sort of self-knowledge as fundamental, and distinct from our knowledge of what we sense, we will not be able to understand what makes the object of self-knowledge a *self*. For to be a self is to be a thinker and an agent, and to be a thinker and an agent is to be capable of a kind of activity that stands in contrast to the passivity of sensation. Nor is this merely a point that must be acknowledged by *theorists* of self-knowledge. Even to be capable of having such mundane thoughts as "I'm in pain," we must have an at-least-implicit conception of the active subject to which sensations belong. For, as I have argued, to understand the use of the first person in ascriptions of sensation requires understanding its connection with our ability to make up our minds through deliberation. It requires, in other words, that our representation of the subject whose sensations are in question imply that this subject also possesses a capacity for spontaneity. And this, in fact, is just what Kant says about "the representation *I*:" it is a representation of "the spontaneity of a thinking subject" (B278). Our conclusion is that only a creature possessed of such a representation can know itself.[32]

[32] For comments on earlier drafts, I am grateful to Niko Kolodny, John McDowell, Kieran Setiya, Susanna Siegel, and Michael Thompson. I am especially indebted to Sebastian Rödl for extensive comments, to an anonymous reader from this Journal for searching criticisms, and to Doug Lavin and Matthias Haase for much help and advice.

References

Ameriks, Karl. 2000. *Kant's Theory of Mind*. Second Edition. Oxford: Oxford University Press.

Anscombe, G. E. M. 1963. *Intention*. Second Edition. Oxford: Basil Blackwell.

—— 1975. "The First Person." In *Mind and Language*, ed. Samuel Guttenplan. Oxford: Oxford University Press.

Armstrong, D. M. 1968. *A Materialist Theory of the Mind*. New York: Humanities Press.

Bar-On, Dorit. 2004. *Speaking My Mind*. Oxford: Oxford University Press.

Bar-On, Dorit and Long, Douglas C. 2001. "Avowals and First-Person Privilege." *Philosophy and Phenomenological Research* 62: 311–335.

Bilgrami, Akeel. 1998. "Self-Knowledge and Resentment." In Wright, Smith and MacDonald 1998.

Brook, Andrew. 1994. *Kant and the Mind*. Cambridge: Cambridge University Press.

Burge, Tyler. 1996. "Our Entitlement to Self-Knowledge." *Proceedings of the Aristotelian Society* 96.

Burge, Tyler. 1998. "Reason and the First Person." In Wright, Smith and MacDonald 1998.

Davidson, Donald. 1984. "First Person Authority." *Dialectica* 38: 101–11.

Edgley, Roy. 1969. *Reason in Theory and Practice*. London: Hutchison.

Eilan, Naomi and Roessler, Johannes, eds. 2003. *Agency and Self-Awareness*. Oxford: Oxford University Press.

Evans, Gareth. 1982. *The Varieties of Reference*. Oxford: Oxford University Press.

Falvey, Kevin. 2000. "The Basis of First-Person Authority." *Philosophical Topics* 28: 69–99.

Finkelstein, David H. 2003. *Expression and the Inner*. Cambridge: Harvard University Press.

Hamilton, Andy. 1998. "The Authority of Avowals and the Concept of Belief." *European Journal of Philosophy* 8: 20–39.

Jacobsen, Rockney. 1996. "Wittgenstein on Self-Knowledge and Self-Expression." *Philosophical Quarterly* 46: 12–30.

Kant, Immanuel. 1902. *Gesammelte Schriften*. 29 vols. Ed. Royal Prussian Academy of Sciences. Berlin: Georg Reimer.

—— 1929. *Critique of Pure Reason*. Trans. Norman Kemp Smith. New York: St. Martin's Press.

Keller, Pierre. 1998. *Kant and the Demands of Self-Consciousness*. Cambridge: Cambridge University Press.

Kitcher, Patricia. 1990. *Kant's Transcendental Psychology*. Oxford: Oxford University Press.
Lewis, David. 1972. "Psychophysical and Theoretical Identifications." *Australasian Journal of Philosophy,* 50(3): 249–258.
Lycan, William G. 1998. "Consciousness as Internal Monitoring." In *The Nature of Consciousness*, ed. Ned Block, Owen Flanagan and Güven Güzeldere. Cambridge: MIT Press.
Moran, Richard. 2001. *Authority and Estrangement*. Princeton: Princeton University Press.
—— 2003. "Responses to O'Brien and Shoemaker." *European Journal of Philosophy* 11: 402–419.
Nichols, Shaun and Stich, Stephen P. 2003. *Mindreading*. Oxford: Clarendon Press.
O'Brien, Lucy. 2003. "Moran on Agency and Self-Knowledge." *European Journal of Philosophy* 11: 375–390.
Perry, John. 1979. "The Problem of the Essential Indexical." *Noûs* 13: 3–21.
Powell, C. Thomas. 1990. *Kant's Theory of Self-Consciousness*. Oxford: Clarendon Press.
Rödl, Sebastian. Forthcoming. *Self-Consciousness*. Cambridge: Harvard University Press.
Ryle, Gilbert. 1949. *The Concept of Mind*. New York: Barnes and Noble Books.
Shah, Nishi and Velleman, J. David. 2005. "Doxastic Deliberation." *Philosophical Review* 14(4): 497–534.
Shoemaker, Sydney. 1963. *Self-Knowledge and Self-Identity*. Ithaca: Cornell University Press.
—— 1988. "On Knowing One's Own Mind." *Philosophical Perspectives* 2: 183–209.
—— 1990. "First-Person Access." *Philosophical Perspectives* 4: 187–214.
Wilson, George M. 2004. "Comments on *Authority and Estrangement*." *Philosophy and Phenomenological Research*, LXIX: 440–447.
Wittgenstein, Ludwig. 1953. *Philosophical Investigations*. Trans. G. E. M. Anscombe. New York: Macmillan Publishing Company.
Wright, Crispin. 1987. "On Making Up One's Mind: Wittgenstein on Intention." In *Logic, Philosophy of Science and Epistemology: Proceedings of the 11th International Wittgenstein Symposium*, ed. Paul Weingartner and Gerhard Schurz. Vienna: Hölder-Pichler-Tempsky.
—— 1991. "Wittgenstein's Later Philosophy of Mind: Sensation, Privacy and Intention." In *Meaning Scepticism*, ed. Klaus Puhl. New York: De Gruyter.

—— 1998. "Self-Knowledge: The Wittgenstein Legacy." In Wright, Smith and MacDonald 1998.
Wright, Crispin, Smith, Barry C. and MacDonald, Cynthia, eds. 1998. *Knowing Our Own Minds*. Oxford: Clarendon Press.