PHILOSOPHICAL TOPICS
VOL. 28, NO. 2, FALL 2000

The Basis of First-Person Authority

Kevin Falvey
University of California, Santa Barbara

I

Recently a colleague stopped by my office and asked what I thought about the latest draft of a student's dissertation. I told him that I believed that it was ready to be approved. He asked me why I thought so, we discussed the matter a bit, and he went away knowing what I thought about the student's work. Here we have a characteristic reliance on *first-person authority*: wishing to know what I thought about the dissertation, my colleague simply asked me, and took my response at face value. My concern in this paper is the nature of the authority each person is recognized as having on matters having to do with his own mind. Consideration of our practices surrounding encounters such as this suggests the following as some of the essential features of first-person authority.

First, it is normally assumed to hold over a wide range of mental states, including beliefs, desires, intentions, emotions, moods, and sensations. In general, if one wants to know what another person thinks, wants, or feels, the best and simplest way of finding out is to ask her. In this paper, I will confine my attention to the authority one has with respect to one's own beliefs, though I believe the account I present will have some application to other propositional attitudes as well.

Second, the authority we are concerned with appears to be genuinely *epistemic*. First-person attributions of belief and other mental states are generally

accepted because individuals are held to *know* what they think and feel. The hearer does not typically treat the speaker's avowals as evidence from which one may draw conclusions about the speaker's mind, in the way that I might infer that my friend is depressed from her frequent disparaging remarks about herself. In a normal avowal, the speaker *informs* the hearer what she thinks or feels.

Third, and despite the feature just mentioned, it is generally not appropriate to ask the speaker for the reasons or evidence on which his claim about his own mind is based. In the example above, when I told my colleague that I thought the student's dissertation was ready to be approved, and he asked me why I thought so, I assumed, correctly, that he meant, "Why do you think the dissertation is ready?" not "Why do you think that you *believe* the dissertation is ready?" Few people would know how to answer the latter question, and it would probably be regarded as an aggressive or presumptuous challenge. Of course, we often accept the word of others on various and sundry topics without question, and it is probable that we have a general epistemic entitlement to do so.[1] But it is not uncommon, and frequently entirely appropriate, to ask another how he knows what he claims about matters of fact in general. I assume that the singular inappropriateness of this question in the face of a person's avowal of his belief indicates that we have a distinctive warrant for accepting such avowals as true.[2]

Finally, first-person authority is definitely not absolute. Apart from simple lying, there are a variety of kinds of self-deception to which persons are occasionally subject, giving rise to situations in which others may know that one's avowals should not be taken at face value. However, while some failures of self-knowledge are more serious than others, the presumption of first-person authority is such that where a person's avowals are not true, this is usually due to some culpable failure on the part of the speaker.[3]

Accounts of the basis of first-person authority in the philosophical canon can be grouped roughly into two classes. The first group, in keeping with the predominant individualism of much of western epistemology and philosophy of mind, sees the intrinsic credibility of avowals as due to their being the product of a special mode of awareness—typically conceived of as analogous in various respects to perception—that each person has of the contents of his own mind. This type of account has come in for much criticism in recent years, and I will have some critical things to say about the perceptual model later in this paper. Here I want simply to note that, in contrast to this type of account, I regard first-person authority as an interpersonal phenomenon, consisting first and foremost in the fact that each of us normally takes the avowals of others at face value. This raises the philosophical question, what is the nature of our warrant for doing so? The answer I will propose makes essential appeal to the social role of avowals

in our practices surrounding the transmission of knowledge through testimony. Indeed, I intend to invert the traditional order of explanation, and will be drawing conclusions concerning the nature of the *speaker*'s knowledge of his own attitudes from a preliminary account of the warrant the *hearer* has for accepting the speaker's avowals as true.

The other main approach to the problem of first-person authority was prominent in mid-twentieth-century philosophy of mind, being largely inspired by Wittgenstein's remarks about verbal expressions of pain, especially his well-known claim that "the verbal expression of pain replaces crying and does not describe it."[4] The leading idea of this approach is that first-person authority is a consequence of the fact that avowals are *expressions* of the speaker's mental states. There are several points of contact between this sort of view and the account to be proposed in this paper. This view, like mine, is rooted in a conception of the role of avowals in communication. Moreover, the idea that avowals of belief are expressions of belief, understood against an appropriate background, will play a major role in this paper. But I want to disassociate myself from the non-cognitivist cast that the expressivist view tended to take on in the hands of some of its proponents. At work here was a supposed contrast between *expressive* and *assertive* uses of language, a contrast familiar from non-cognitivist ethical theories. Taken to its extreme, this involves denying that avowals of mental states are capable of being true or false or of manifesting the subject's knowledge of his own mind. This consequence was attractive to some philosophers who, inclined toward behaviorism, seem to have thought that to allow that avowals purport to state facts requires the postulation of an inner realm of facts for the true ones to correspond to, along with some mysterious cognitive faculty that affords each individual "privileged access" to his own mind. The expressivist account of avowals, in its non-cognitivist guise, held out the promise of an alternative, deflationary construal of the fact (if it is a fact, which is unlikely) that a person can't be mistaken when she avows her mental condition. If the utterance of "I am in pain" is not an assertion but an expression of pain essentially similar to a groan, then it is not the sort of thing to which the notion of a *mistake* applies. J. J. C. Smart seems to be applying this sort of view to something resembling a propositional attitude when he suggests that "saying 'I love you' is just part of the behavior which is the exercise of the disposition of loving someone."[5]

Now to be sure, saying "I love you" in the appropriate context can indeed be an expression of love, as much as a caress or a personal gift. But how is it supposed to follow that the utterance of this sentence is not an assertion, which is true if and only if the speaker loves the person to whom it is addressed? On the face of it, the evidence that such an avowal is an assertion capable of being evaluated for truth is impressive. The sentence

used in making it is syntactically and semantically structured, and appears to be used to attribute to oneself the same property that one could attribute to another using the sentence, 'She loves you,' the truth aptness of which is beyond question. Moreover, I might attempt to comfort someone who doubts the loyalty of her friends by saying, "But all your friends love you," offering in support of this the argument, "I love you, and Sheila loves you, and Sheila and I are your only friends." While this might be cold comfort, it certainly looks like a valid argument. But this cannot be so unless the first premise has the semantic structure and truth aptness it appears to have, and is indeed being presented as true.[6]

One route to the non-cognitivist conclusion might allow that an assertion that *p* may be said to express the belief that *p*, and then claim that a speech-act type cannot express more than one type of mental state. It would follow that regarding an utterance of "I love you" as an expression of love would preclude viewing it as an assertion. But the idea that speech acts can or must be paired off one-to-one with the types of mental states they express is dubious, since cognitive and emotional states commonly occur in clusters, and exhibit systematic connections. Surely, the cry "It's raining!" coming at the end of a long drought, can express both surprise and delight, as well as the speaker's perceptual knowledge that explains the surprise and justifies the delight.

The thesis that avowals are expressive is often paired, in Wittgenstein's writings and elsewhere, with the denial that they are *reports* of inner states.[7] This seems to me to be correct, at least for an important class of avowals of propositional attitudes.[8] But it would be a mistake to conclude that the sentences used in such avowals are not truth apt from the premise that they are not reports.[9] 'Report' is a term of linguistic pragmatics; it refers to something one can do or make using an indicative sentence. The notion of reporting is cognate with the notion of observation: reporters are people who are sent off to far-flung places to observe what is happening and report back to us on their findings. However, there are other things one can do with indicative sentences besides make reports, and one should not draw conclusions regarding the syntax or semantics of a sentence uttered—such as that it is not in the indicative mood, or not capable of being true or false, from the pragmatic fact that it is not a report. Performatives, for example, are not reports, but there are reasons for thinking that an utterance in the appropriate context of "I promise to help you with your work tomorrow" is an assertion that is made true by the act of uttering it.

The thesis that certain avowals are not reports should therefore be understood as a denial that perception is an appropriate model in terms of which to think about first-person authority with respect to the mental states associated with them. Pragmatically, some avowals of mental states are

expressions rather than reports. But this entails nothing about the semantics of self-ascriptions of these states. Indeed, if we avoid the errors of non-cognitivist expressivism, and take seriously the idea that avowals have an expressive function, while retaining the natural and compelling idea that the sentences used in making them are capable of truth or falsity, and presented as true—so that such avowals also have the pragmatic status of assertions—then a simple but, I think, powerful truth comes into focus. Let φ be a mentalistic verb phrase such that the utterance "I φ" is expressive of the mental state φ describes. It would seem that if such an expression is *sincere,* then the speaker is in the given mental state. Since this is precisely what the speaker asserts in uttering these words, the sincerity of the utterance suffices for its truth.[10] In other words, perhaps some avowals have the unique property that the gap between sincerity and truth that exists for most assertions collapses here. This would surely go some way toward explaining first-person authority, provided that sincerity can be assumed to be the norm in interpersonal communication.

This is the idea that will be pursued in this paper. In the next section, I outline the background necessary for seeing self-ascriptions of belief as expressions of belief of a distinctive kind. This will lead to a preliminary account of the hearer's warrant for accepting the avowals of another, and then to a proposal concerning the nature of the speaker's knowledge of his own mind that is manifest in sincere avowals. The proposal is elaborated and qualified in sections III and IV. In section V I argue that my account does a better job of accounting for the main features of first-person authority than does the perceptual model, and in section VI I discuss how the account handles certain notable failures of first-person authority. I conclude with some brief remarks on the relation between my account and externalism about mental content.

II

Wittgenstein's expressivist account of avowals of pain was part of an attempt to demystify the avowability of mental states by assimilating first-person ascriptions of such states to natural expressions of pain, anger, fear, and other emotions. Part of the point of this is that we share not only these mental states, but also characteristic ways of expressing them, with other animals. Wittgenstein seems on occasion to have considered extending this idea to the propositional attitudes, in one place finding a "natural expression of intention" in the movements of a cat stalking a bird.[11] However, given the connection between the propositional attitudes and the notion of rationality, I think a more plausible background against which to develop an expressivist

account of avowals of the attitudes is the distinctively human, social practices that involve asking for, giving, and using reasons for thought and action. In the case of belief, we need to look at the role that avowals play in the transmission of knowledge through testimony.

Idealizing somewhat, it seems reasonable to say that the point of the language game in which assertion finds its home is the transmission of knowledge. Since knowledge cannot be transmitted unless it is possessed, we are led naturally to the idea that the assertion that *p* is warranted if and only if the speaker knows that *p*. This idea may be unfamiliar; why shouldn't justified belief be sufficient for warranted assertion? To begin with, note the relevance of such challenges to an assertion as, "You don't know that!" The fact that the speaker to whom this is addressed does not know what he has just asserted does not call into question the propriety of his assertion unless knowledge is required for warranted assertion. In addition, consider the deviant behavior of Jones in the following dialogue:

Smith: Where is Susie?
Jones: She's at the playground.
Smith (into his cell phone): It's OK. Jones knows where Susie is.
Jones (overhearing): Wait, I didn't say I *know* she's at the playground; I just said she's at the playground.

Jones's last remark is certainly odd, and its oddness is understandable if knowledge is required for warranted assertion. For then if Jones does not know that Susie is at the playground, he shouldn't state flatly that she is; he should say instead that he thinks she is.

It might be suggested that it is merely an implicature of the assertion that *p* that the speaker knows that *p*. The oddness of Jones's last remark above would then be glossed as a deliberate flouting of this implicature. The case would thus be assimilated to a case in which a person answers a question with a disjunctive assertion, and then reports later that he knew all along which of the disjuncts was true. But there is a difference between these two cases. It cannot reasonably be suggested that the maxim "say as much as you know" is as fundamental to the nature of assertion as the rule: "say *only* what you know." That the point of the game of asserting is the transmission of knowledge makes the latter the fundamental rule governing moves in this game.

Perhaps this case shows only that for an assertion that *p* to be warranted the speaker must *think* he knows that *p*. But once it is recognized that knowledge plays some crucial part in the warrant for an assertion, simplicity argues in favor of taking the fundamental rule to be: assert that *p* only if you know that *p*. Obviously, in the application of any rule, the person applying the rule must rely on his best judgment as to what the rule entitles her to do in a given situation. One may reasonably believe that an action is warranted, when in fact it is not. We may apply here the distinction made in ethics between a justification and an excuse, and say that a person who asserts that *p* because he reasonably but falsely thinks he knows that *p* should be excused for making an assertion that was in fact unwarranted.[12]

Given that one is warranted in asserting that *p* if and only if he knows that *p*, a simple question such as "Where is the library?" presupposes that the person to whom it is addressed knows the correct answer to it. (More cautious questions that do not carry this presupposition include, "Do you know where the library is?" and "Do you think that the library is over there?"). Among the relevant responses to a simple question are simple answers to it, such as "two blocks down on the right," as well as the response, "I don't know": a simple rejection of the question, which points out that it is based on a false presupposition. In addition to these responses, it is sometimes in order to respond to the simple question, "Where is the library?" by saying,

(i) I believe the library is two blocks down on the right.

This is neither a simple answer to the question nor a simple rejection of it, though it shares some features in common with each of these types of response. Like a simple rejection of the question, it undermines the question's presupposition, since part of the point of this use of the words 'I believe' is to indicate that the speaker does not know, or doesn't think she knows, the correct answer to the question. But like a simple answer to the question, responding in this way is intended to provide potentially useful information concerning the topic of the question. Semantically, the assertion of (i) is a statement about the mind of the speaker; in uttering it the speaker attributes to herself the same belief she could attribute to her friend by asserting,

(ii) My friend believes the library is two blocks down on the right.

But the mind of the speaker who utters (i) is relevant in this context only to the extent that it sheds light on the location of the library. In asserting (i) the speaker is *endorsing* the truth of the proposition that the library is two blocks down on the right. Her endorsement is tentative or provisional, but her assertion shares with the unqualified assertion,

(iii) The library is two blocks down on the right

the feature that (i) expresses a *commitment* to the claim that (iii) is the correct answer to the question, or at least to the claim that this is more likely than other answers to be correct. The person who asserts (i) makes a *qualified assertion* that the library is two blocks down on the right.[13] This pragmatic similarity between (i) and (iii) is not shared by (ii). My assertion of (ii) does not commit me, even in a qualified way, to any claim about the location of the library.[14]

The commissive aspect of both (i) and (iii) is most easily seen by noting that both (iv) and (v) are deviant or defective assertions:

(iv) The library is two blocks down on the right, but it isn't two blocks down on the right

(v) I believe the library is two blocks down on the right, but it isn't two blocks down on the right.

The assertion of (iv) is defective because, being a contradiction, it cannot possibly be true, and so cannot convey any useful information about the location of the library. Although an utterance of the "Moore-paradoxical" (v) could be true, it is also useless as an assertion, which is not mysterious if it is born in mind that the point of uttering the first conjunct of (v)—like the point of uttering the first conjunct of (iv)—is to express an epistemic commitment as to the location of the library. In uttering (v), the speaker denies in the second conjunct the very proposition he expressed a commitment to in the first conjunct. Unlike (iv), (v) describes a possible state of affairs, one in which the speaker has a false belief about the location of the library. But if this state of affairs is actual then the speaker is obviously not a useful source of information on the location of the library. Hence, (v) can no more be used to convey information on this topic than (iv).[15]

To summarize: (i) is pragmatically like (iii), and unlike (ii), in that both (i) and (iii) express, at least in a qualified way, commitment to the truth of a proposition about the location of the library. On the other hand, (i) is semantically like (ii), and unlike (iii), in that each of (i) and (ii) ascribes a belief to a particular individual, so that (i) and (ii) are true provided the individual referred to has that belief, regardless of the actual location of the library. Given that the assertion of (iii) is warranted if and only if the speaker knows that the library is two blocks down on the right, we can say that the pragmatic role of an assertion of this type is to express the knowledge that that is where the library is located.[16] Furthermore, since the assertion of (i) shares with the assertion of (iii) the pragmatic property of expressing a commitment to the same proposition about the library, the only essential *pragmatic* difference between the two assertions is that the former is qualified by the implicature it carries to the effect that the speaker lacks knowledge. It is thus natural to say that the role of the assertion of (iii) is to express the belief that the library is two blocks down on the right.

Suppose we now introduce the expressivist idea mentioned at the end of the last section and say that an individual who expresses the belief that *p* does indeed believe that *p* provided his expression of belief is *sincere*. In addition, since (i) has the semantic role of ascribing to the speaker the belief that the library is two blocks down on the right, the assertion of (i) will be true provided the speaker believes this. It follows that if the speaker asserts (i) sincerely, his assertion will be true. The Janus-faced character of self-

ascriptions of belief, the fact that they play the role of *expressing* the very mental state that they simultaneously *ascribe* to the speaker, ensures that the gap that typically exists between the sincerity and the truth of an assertion does not exist here. This is the basis, I think, of first-person authority with respect to avowals of belief. As a first approximation, I suggest the following formulation:

(E1) An individual is warranted in accepting as true another person's avowal of belief, unless there is reason to think the avowal is insincere.

The principle explains why it is normally appropriate to accept the avowals of others without question. Provided the avowal of a belief is sincere, which may be assumed to be the norm in interpersonal communication, the speaker is giving direct expression to, making manifest, what he believes.

Consider now a normal conversational interchange in which the speaker sincerely avows a belief, and the hearer accepts it without question. The hearer is warranted in doing so, there being no reason to think the speaker was not sincere. The hearer goes away with a warranted true belief, which surely amounts to knowledge, concerning what the speaker thinks.[17] This is a special case of the acquisition of knowledge through testimony. As such, the hearer cannot acquire knowledge of the fact stated by the speaker —that the speaker has a certain belief—unless the speaker himself knows this to be true. (Again, we are not concerned here with a situation of the kind in which I come to know that my friend is depressed from her disparaging remarks about herself. In that kind of case I learn from her remarks something about her mind that she may not know. In contrast, in a sincere avowal the speaker informs the hearer what the speaker thinks.) What is the nature of the speaker's warrant for his assertion? What sort of knowledge is manifest in a sincere avowal? Well, it appears that in sincerely expressing what he believes, the speaker's utterance is automatically true. What more is relevant to the speaker's *warrant* for the knowledge that, by hypothesis, his self-attribution manifests? For a first approximation, I suggest, as a kind of null hypothesis, that the answer is: nothing. That is, I suggest that the following epistemic principle holds:

(E2) A speaker's avowal of belief is warranted provided it is sincere.

Similarly, I suggest that the self-knowledge that is manifest in paradigmatic exercises of first-person authority consists in the capacity a normal individual has to give expression to what he believes. I emphasize, however, that the principles (E1) and (E2) are first approximations only. They will be qualified later on (in section IV).

III

At this point a certain type of "internalist" epistemologist might raise the following objection. I am claiming that first-person authority consists in the capacity to give expression to what one believes, but I am determined to avoid the shortcomings of older expressivist accounts. The capacity in question here must be understood as a rational, cognitive capacity, and some account must be given of what constitutes a responsible exercise of this capacity. Now, even granting that the sincerity of the avowal, "I believe that *p*," suffices for its truth, since sincerity in such an assertion consists in saying what one believes, the speaker must know that he believes that *p* in order to know what would *count* as a sincere expression of his belief. Unless his assertion is the product of such knowledge, the expression of his belief can hardly be described as a rational exercise of cognitive faculties. Even if his avowal were sincere, and hence true, he could not be held responsible for it if it were not the expression of knowledge. Therefore, the principle suggested at the end of the last section cannot constitute the *foundation* of first-person authority from the speaker's standpoint. Some more basic account of how the speaker knows what she believes must be given, perhaps along the lines suggested by the perceptual model.

I agree that avowals must involve the rational exercise of cognitive faculties—they must be, in Sellars's phrase, moves within "the space of reasons"—if the capacity of an individual to express his beliefs is to be seen as constitutive of the kind of self-knowledge that is manifest in first-person authority. But I think this objection underestimates the resources available to the expressivist account I am proposing to provide what is needed here. As a start on dealing with the objection, I want to question the objector's claim that *sincerity in an assertion consists in saying what one believes*. In particular, I think it is not always the case that, when a person asserts that *p* (with or without qualification[18]) her assertion is sincere if and only if she believes that *p*.[19]

To see this, consider the following example. Suppose that Jill is about to drive from Santa Barbara to Los Angeles, and she asks me whether it would be faster to take Highway 1 or Highway 101. I believe that the latter route would be faster, but for some reason, I wish to delay her arrival in L.A. Suppose also that as it happens, I believe that she thinks that I am a fool, and has asked for my advice with the intention of doing precisely the opposite of whatever I suggest. With all of this in mind, I tell her that taking 101 would be faster, because this seems to be the best way of ensuring that her arrival in L.A. is delayed. Now, it could hardly be said that this was a sincere expression of my belief about the driving conditions. Although I said, "101 would be faster," and this is what I believe, I said this with the intention of deceiving her as to the truth, as I saw it.[20] Contrast this case with the following variant. Here again, Jill asks which route from Santa Barbara to L.A. is faster, and I tell her that 101 is faster, but this time I say this because I believe that it is the correct answer to her question. We may still assume, if we wish, that I want her arrival in L.A. to be delayed, and I believe that she thinks I am a fool. But this time this desire and belief play no role in my decision as to how to answer her question. It may cross my mind that as a result of my response she will take Highway 1 and be delayed, but I think, "If she wants to go against my advice that's her business. She asked me what I thought and I told her." In this case I am the sort of person who "says what he thinks" and lets the chips fall where they may. It is this kind of "saying what one thinks" that is the hallmark, I think, of sincerity.[21]

In the second example I treated the question, "How should I respond to Jill's question?" as reducing to the question, "Which do I think would be the best route?" and I responded accordingly. In the first example this was not the case. Having answered the question about what I thought, there remained for me the further question, "Now what should I tell her?" and considerations additional to what I thought about the driving conditions were appealed to in answering this question. Whenever the question of what to say to an interlocutor is separated in this way from the question of what I think, whatever I eventually say in response will be tainted with insincerity. It will reflect, in part, some motive other than the desire to convey what I think. Even if I do end up saying what I think, this will only be a coincidence, as in the first example.[22] The calculation that lies behind my response means that what I say is not a simple and straightforward *expression* of my belief. Sincerity in an assertion, then, consists in saying what you think *because* that is what you think. It should be characterized negatively, as the lack of pretense, calculation, or ulterior motives for the assertion.[23]

To return to the objection that began this section, it relies on another claim that deserves scrutiny, namely, that "the speaker must know that he believes that *p* in order to know what would count as a sincere expression of his belief." This is misleading at best, and false if it is understood to imply that the speaker must know what he thinks *before* he can be in a position to express it. For it is a consequence of the fact that "I believe that *p*" expresses qualified commitment to the truth of *p* that one should go about answering the question, "Do I believe that *p*?" by attempting to answer the question, "Is it true that *p*?" One does not typically determine what one believes by looking into one's soul. As Gareth Evans has observed,

> [In] making a self-ascription of belief, one's eyes are, so to speak, and occasionally literally, directed outward—upon the world. If someone asks me "do you think there is going to be a

> third world war?" I must attend, in answering him, to the same outward phenomena as I would attend to if I were answering the question "will there be a third world war?"[24]

This is a consequence of the expressive role of avowals of belief because, in considering how to respond to the question, "Do you think there will be a third world war?" one is considering whether to undertake a commitment to the truth of the proposition that there will be a third world war, and such a commitment should be undertaken just in case the available evidence "out there" in the world indicates that this proposition is true or probably true. Similarly, and generalizing somewhat, if I am asked, "What do you think of McCarthy's new novel?" I must turn my attention to the novel, not to my own mind. The question of what I think about X should be addressed by considering what I *ought* to think about X, where this question in turn amounts to the question, "What is the truth about X?" or more cautiously, "What judgment about X is best supported by the available evidence?"" Questions about what I think are in this way "transparent" to corresponding questions about the world.[25]

The transparency of questions about what I think to corresponding questions about the world is most clearly on display when one is asked what one thinks about some matter that has not been previously considered. "What do you think the weather is going to do?" a friend asks, whereupon I look at the sky and say, "I think it's going to rain." There is no question here of first determining—in the sense of *discovering*—what I think and then finding words to express it. On the contrary, I determine—in the sense of *making it true*—that I think it's going to rain in undertaking the qualified commitment to the truth of the proposition that it is going to rain that is expressed as it is undertaken. On the other hand, what about situations in which one is asked for one's opinion about some matter that one has previously considered, and about which one has a fully-formed belief? Here it might be thought that Evans's transparency fails to hold, and that in such cases I do look into myself, to my memory in particular, for the answer to the question about what I think. If this is so, then Evans's transparency is of limited interest in connection with the phenomenon of first-person authority, since most of our avowals are of standing beliefs, and such avowals are just as authoritative as those that express the act of making up one's mind.[26] I think, however, that it is a mistake to view the situation in this way. Evans's point holds even with respect to avowals of standing beliefs, provided the role of memory in such avowals is properly understood.

Consider again the example with which I began this paper. My colleague asks me what I think of the student's dissertation, and having read it several days before, I say that I think it is ready to be approved. My response relies on memory, but I do not think that memory plays a *justificatory* role

relative to the judgment, "I think it's ready." That is, memory does not provide reasons or evidence for thinking *that I think* the dissertation is ready. Instead, memory presents as true the content previously endorsed—the dissertation is (probably) ready to be approved—together with the propositions about the work that formed the basis for that endorsement. Memory preserves these judgments, and the justificatory relations among them, making them available for re-endorsement at later times. To cite one's memory as a reason for thinking that one thinks that *p* would be odd in roughly the same way that the following would be odd. A student comes into my office with a copy of one of my publications, points to a certain claim that occurs in it, and asks, "Do you really believe that?" I say in response, "Well, that is indeed my paper, and I don't recall having changed my mind about anything in it, so I guess I do believe that." Here I am treating the fact that I wrote this sentence as *evidence* that I have a certain belief. Such an attitude is perfectly appropriate when attributing a belief to another, but to self-ascribe a belief on such a basis would surely be irrational, precisely because I would not be treating the question of whether I had the belief as transparent to the question whether it was true. The crucial point here can be expressed in terms of the contrast between looking inward and looking outward that figures in Evans's remarks about the transparency of avowals. Memory, when it is relied on properly in avowing a standing belief, should be thought of as a faculty that expands the range of "outward" phenomena one is in a position to make a judgment about. In the dissertation example, my present judgment, "I think the dissertation is ready," is the product of my "looking," via memory, at *the dissertation*, and endorsing (with qualification) a proposition about it, self-attributing the belief I express at the same time. My judgment is still sensitive to the relevant facts about the student's work, as can be seen from the fact that I always have the option of withdrawing my commitment to the truth of the propositions memory presents as true, if new facts have come to light, by saying, for example, "I thought it was ready to be approved, but now I am not so sure." So even where memory is relied on in avowing a belief, the question of what I think about X is still transparent to the question, what are the facts about X?[27]

I am now in a position to complete my response to the objection I raised at the beginning of this section. Doing justice to the idea that sincere avowals of belief involve the rational exercise of cognitive faculties does not require the postulation of a special cognitive faculty that informs each individual what she believes. Sincerity in response to the question, "Do you think that *p*?" (or alternatively, "What do you think about X?") consists in the subject's treating the question of what to say in response as reducing to the question, "Do I believe that *p*?" (alternatively, "What do I think about X?"). And a rational individual will treat this question as transparent to the

question, "Is it true that *p*?" (or, "What are the facts about X?"). Thus a sincere and rational response to the original question, whether memory is relied on or not, will take the form of an appropriate judgment about the subject matter in question. The speaker may choose to prefix her judgment with the words *I think,* which has the pragmatic effect of qualifying her endorsement of the proposition the truth of which she is committing herself to, and the semantic consequence that her commitment is expressed by a sentence that means that she has the belief she expresses, so that, this being the case in virtue of her sincerity, what she asserts is true. Because of the Janus-faced character of avowals of belief, their status as rational exercises of cognitive faculties, under their self-attribution aspect, is inherited from the way in which such faculties are deployed in avowals, under their endorsing aspect. Nothing more is required in order to knowledgeably self-ascribe the belief that *p* than the conceptual capacity and minimal rationality requisite for responsibly judging that *p*.[28]

IV

Now it might be objected that while one generally treats the question whether one believes that *p* as transparent to the question whether it is true that *p*, it is possible to distinguish between the question what I do in fact believe, on the one hand, and what would make the belief true, on the other. After all, as noted above in connection with Moore's paradox, it might be true on a given occasion that although I believe that *p*, it is not true that *p*, or vice versa. To say that I do not ordinarily distinguish between these two questions is to say that ordinarily, I treat the question of what I believe as a question about what *to* believe, where this in turn is transparent to the question as to the truth concerning the world. However, there would seem to be circumstances in which one asks oneself, "What do I think about X?" not under the guise of deliberating concerning what one ought to think about X, but rather, under the assumption that one already has a determinate belief about X, but does not know what it is.[29] Such is perhaps the position of the stereotypical subject in psychoanalysis who asks, "What do I really think about my father?" It seems to me that this situation is rarer than one might think. In most cases, wondering what one thinks amounts to wondering what *to* think. But let us grant that there are some situations where one seeks to determine what one currently believes, while "bracketing" the question of the truth of the matter under consideration. The judgment, "I believe that *p*," reporting the results of such an investigation will not, in such a case, be an *expression* of my belief that *p*, but a *report* of an introspective observation; a *description* of a portion of my inner world. Perhaps the stereotypical psychoanalytic situation is one such. Indeed, in such a context it might even be possible for one to utter an intelligible Moore-like utterance, at the moment when the scales fall from one's eyes: "Now I see it! I believe my father was cruel to me, but he wasn't cruel to me!" We need to distinguish, then, between judgments of the form "I believe that *p*" that are intended as qualified endorsements of the embedded proposition, and judgments of the same form that report the results of introspection. Borrowing Evans's terminology, let us call the former *transparent* avowals, and the latter *introspective* avowals.[30] A sentence of the form 'I believe that *p*' now appears to have two uses. It can be used as a way of making a qualified assertion that *p*, or to report the result of an episode of introspection. This is not to say that we have here two distinct senses, however. The sense of 'believe that *p*' is univocal across these two uses. Since the pragmatics of these utterances is crucial, however, the epistemic principles (E1) and (E2) stated at the end of section II must be reformulated as follows:

(E1') An individual is warranted in accepting as true another person's avowal of belief, unless there is reason to think that the avowal is insincere or not transparent.

(E2') A speaker's avowal of belief is warranted provided it is sincere and transparent.

Having distinguished between transparent and introspective self-ascriptions of belief, I want to emphasize that the transparent judgments remain fundamental. They are so in at least four respects. First, the vast majority of our judgments of the form, "I believe that *p*," or the more colloquial form, "I think that *p*" are cognitively transparent judgments which are put forth without the slightest bit of introspection. "I think it's going to rain," "I think I left my wallet in my other coat," "I think Bush will emerge victorious"—the contexts in which such assertions are ordinarily made have little to do with introspection. My attention in making such judgments is not upon my own mind but upon the world—on the weather, on the contents of my pockets, or on the contents of the morning paper. They are ways of asserting, respectively, that it's going to rain, that I left my wallet in my other coat, and that Bush will emerge victorious, where in each case, the force of the claim made is qualified by the prefix "I think." Gilbert Ryle rightly drew attention to the centrality of such judgments—he called them "unstudied utterances"—in any account of self-knowledge.[31] I differ from Ryle, however, in that I think they reveal a deep asymmetry between our knowledge of our own minds on the one hand, and our knowledge of other minds and the world in general, on the other.[32] The intrinsic credibility of such utterances stems from the way in which their pragmatic features conspire with their

truth conditions in such a way that they are true, provided they are sincere. Sincere, transparent avowals are expressive of the very mental states they self-attribute.

The second respect in which cognitively transparent judgments are fundamental is revealed by the fact that the epistemic situation of the subject of the stereotypical psychoanalytic judgment is inherently unstable. The instability arises because in making a non-transparent judgment, one temporarily brackets the question of the truth of the content of the belief one attributes to oneself. This stance can only be adopted temporarily, for to persist in it would amount to an attempt to disavow the commitment to the truth of the believed proposition that is of the essence of belief. Such an attitude is deeply incoherent. For this reason, even in a situation where, as a result of introspection, I discover that I believe that *p* and state my discovery in the form of a genuinely introspective judgment to the effect that I believe that *p*, the belief I have just discovered can only survive its discovery if I am prepared to take up and endorse its content in the form of a cognitively transparent judgment to the effect that I believe that *p*.[33]

Third, it is one's transparent avowals that embody, first and foremost, the first-person point of view. The epistemic asymmetry that exists in respect of my knowledge of what I believe on the one hand, and my knowledge of what Smith believes on the other, is mirrored in the pragmatic asymmetry that exists in respect of the judgments, "I believe that *p*," and "Smith believes that *p*." When it is made transparently, the former judgment, but not the latter, commits me to the truth of the embedded proposition. The authority of the former judgment stems from *my* authority, as a rational subject, to make up my own mind on the question whether to believe that *p*. When I seek to determine by introspection whether I believe that *p*, bracketing the question of the truth of the proposition that *p*, I am in effect abandoning the first-person perspective that accounts for the authority of my first-person judgments, and adopting an essentially third-person stance toward myself. For certain purposes, of course, the adoption of such a critical, objective point of view on my own mind may be precisely what is required, and there are surely some kinds of self-knowledge that can only be acquired through the adoption of this point of view. But they are not, I think, the most basic kinds. The most basic kind of self-knowledge is the knowledge that is manifest in one's ability to deliberate responsibly concerning what to believe. For this reason, it is accessible only from and through the first-person perspective from which one carries out such deliberations. Our genuinely introspective judgments are typically based on inference and carry no special authority.[34] In this domain, Ryle reigns: in principle, the means by which I arrive at such judgments are the same as those I employ in attributing a belief to someone else. When I ask, in the introspective way, whether I really believe that my father was cruel to me, I rely in a properly justificatory way on my memories of what I have said and done, and perhaps on imagination concerning how I might behave in various possible circumstances. *Mutatis mutandis*, these are just the same sorts of things I rely on in attempting to answer the question whether Smith believes that his father was cruel to him.

Finally, the need to adopt the objective, third-person stance mentioned in the last paragraph, and go in for a bit of soul-searching in an attempt to discover what one really thinks, is usually occasioned by the realization that some of one's prior avowals have not been entirely sincere, transparent, or rational. The introspective question, "Do I really think that my father was cruel to me?" arises when one discovers reasons for thinking that some of one's past claims to this effect were attempts at getting attention or instilling guilt in one's father, rather than attempts at stating truths. A person who was always ideally sincere and rational would not, I think, need to question himself in this way. That there are no such people does not undermine the fact that treating the question, "Do I believe that *p*?" as transparent to the question, "Is it true that *p*?" is the relevant normative ideal. Our avowals are authoritative to the extent that we approximate to this ideal.[35]

V

The most frequently heard criticism of the perceptual model of self-knowledge points out that there is nothing in the phenomenology of knowledge of one's propositional mental states corresponding to the role that perceptual experience plays in the formation of perceptual beliefs. It has become standard, however, to allow that this criticism is not serious, because the perceptual theorist need not maintain that the analogy he is pressing between self-knowledge and perceptual knowledge is that exact. The analogy may be held to consist simply in the idea that just as, for example, the presence of a blue tie in front of a person with normal vision will generally cause him to believe that there is a blue tie there, so the fact that one believes that there is a blue tie there will generally cause one to believe that one believes that there is a blue tie there.[36] I do not think, however, that the perceptual theorist should be let off so easily on this matter. To begin with, the explanatory power of the perceptual model diminishes as the differences between perception and self-knowledge multiply. This observation can be sharpened into a more serious criticism if we recall that the analogy with perception is supposed to illuminate the status of avowals of attitudes as a kind of *knowledge*. It is therefore surely reasonable to hold the perceptual theorist to the constraint that whatever differences he acknowledges between perceptual judgments

and judgments about one's mental states, he must maintain that the nature of the warrant for the knowledge manifest in the two kinds of judgment is the same, at least in its essential respects. Now if he then proceeds to write off the lack of a phenomenologically distinct "inner sense" as an inessential difference between the two kinds of knowledge in question here, the perceptual theorist appears to be committed to the claim that the nature of perceptual experience plays no essential role in an account of the warrant for perceptual judgments. And this is surely an implausible claim. I do not mean to suggest that perceptual judgments are typically made on the basis of inference from premises expressible in terms of how things look or sound to one—I take it as well established that that is not the case. Perceptual experience is ordinarily transparent to the facts it puts one in contact with, so that the presence of a blue tie on the counter will indeed normally cause an observer to believe that this is the case, without any ratiocination on his part. But it is crucial to construing the causation here as a *rational* response to the situation, one that can warrant the normative status of knowledge, that the conditions over which it is appropriate to defer in this way to the deliverances of the senses are limited, that some rough understanding of the limitations is assumed in normally competent observers, and that the content of a perceptual experience *can* be explicitly adverted to in support of a perceptual judgment that has been called into question. Having judged that the tie on the counter is blue, if it is pointed out that the artificial light in here is not conducive to making accurate color judgments, one can say, "Yes, I know, but I took it outside and *it looks blue out there too.*"[37] A full account of the warrant for perceptual judgments must address these normative aspects of the way in which perceptual judgments are made, criticized, and evaluated. The broad perceptual theorist of self-knowledge, on the other hand, must think he can parlay the mere statistical fact that perceptual beliefs usually co-vary with the states of affairs that would make them true into a notion of warrant for these beliefs. This is to ignore all the normative features that justify locating perceptual judgments within the Sellarsian "space of reasons." If statistical reliability is all there is to perceptual knowledge, then one really should attribute such knowledge to thermometers and weather-forecasting bunions.

In fact, persons are held responsible for their avowals to a much greater extent than they are for their perceptual judgments. The physical and biological conditions of the normal functioning of the perceptual apparatus provide for the possibility of erroneous perceptual judgments that are no fault of the subject, such as those due to colorblindness or optical illusions. But as was discussed in section I, erroneous self-ascriptions are generally due to some culpable failure on the part of the subject. If the basis of first-person authority were some mechanism that typically causes the subject to form

true beliefs about what he believes, the reliability of the mechanism would perhaps justify an initial presumption of irrationality in the face of a sincere but false avowal. But it would seem that such a presumption could be overridden by the finding that the subject's second-order belief-forming mechanism had simply malfunctioned. The perceptual model would thus seem to predict that there could be false avowals for which the subject is no more liable to criticism or reproach than would be, say, a person who has made a false color judgment due to colorblindness or a clever optical illusion.

Notice also that in eschewing the idea that beliefs about one's beliefs are based on the representations of inner sense, the broad perceptual model substitutes one unfortunate phenomenology for another. For I do not just find myself inclined, unaccountably, to say that I think, for example, that Bush will be declared victorious. In the normal case, I form this belief as the appropriate response to the available reasons for thinking its content true. There is a kind of rational agency at work here: I am undertaking an epistemic commitment which, on the expressivist account presented above, is directly expressed in my avowal. The perceptual model, in contrast, postulates an intervening causal transaction between the formation of a belief based on a certain set of reasons, and the belief that one has the first belief, where it is this latter, second-order belief that is expressed in a judgment of the form, "I believe that *p*," on this model. This is to represent the subject as a passive spectator of the belief-forming processes within him, and reduces his self-ascriptions to something like the scripted pronouncements of a government spokesperson.[38]

At its best the appeal to a perception-like second-order belief-forming process is otiose, no such intervening step being needed to account for the ability of an individual to avow his beliefs. At its worst, the introduction of such a process severs the warrant for the avowal from its source in the subject's authority, as a rational being, to make up his own mind on the question of what belief is best supported by the evidence. I argued in section III that when one avows a belief on some topic on which one made up one's mind in the past, the subject's memory does not "tell" him that this is what he believes. To rely on memory in this way would amount to an attempt to disavow the commitment to the truth of one's own beliefs that is distinctive of the first-person point of view. This point can be generalized, I think, to apply to whatever mechanism the perceptual theorist of self-knowledge postulates through which first-order beliefs cause second-order beliefs. Aside from the rather special circumstances discussed in section IV, nothing can or should "tell" me what I think, even in a metaphorical and phenomenologically empty sense of "tell."

VI

In an important paper, Paul Boghossian has criticized accounts of first-person authority that depict the knowledge underlying it as "cognitively insubstantial," on the grounds that they represent self-knowledge as too-easily attained.[39] I suspect that some might think that the account of first-person authority presented in this paper deserves to be described as representing it as cognitively insubstantial, though the matter is not entirely clear.[40] I emphasized in section III that avowals of belief are genuinely cognitive acts, although the cognitive faculties involved in making them are just those that enable us to make judgments about the world in general. On the other hand, the absence, in my account, of a distinct faculty or process through which second-order beliefs are formed, together with my central claim, that the sincere, transparent avowals of a rational individual are all true, might lead one to wonder whether I have depicted self-knowledge as too easily attained. And Boghossian is surely right that any account of first-person authority must explain why it is sometimes difficult indeed to know one's own mind. However, in attempting to discharge this latter obligation, I want to call attention to the fact that self-knowledge is frequently as much a *practical* achievement as a cognitive achievement. This accounts, I think, for a good deal of its difficulty. It also provides the best explanation of why false avowals are generally due to some culpable failure on the subject's part.

One dimension of the commitment to the truth of the proposition endorsed in a sincere avowal of belief is a commitment to take the proposition as a premise, where it is relevant, in one's subsequent theoretical and practical deliberations. A person who believes that *p* intends to act as though it were true that *p*, where what counts as so acting will depend on her other beliefs, desires, and intentions. This is the point at which the speaker's warrant for her avowals meshes with the warrant another person has for attributing the avowed attitudes to her. Third-person belief attributions are warranted by the available behavioral evidence, including of course the subject's verbal behavior, as well as whatever evidence may be available concerning how the subject can be expected to behave in the future. Seen from this angle, the presumption of first-person authority amounts to a concession that persons normally have the self-control to ensure that their actions will be appropriate to their words. For if an avowal of belief is sincere and transparent, then the speaker believes what she claims to, and hence intends to act in such a way that her subsequent behavior can be expected to continue to support the attribution of the belief to her, provided she has the self-control necessary to carry out her intentions. Hence, under normal circumstances, a speaker's word suffices for another to attribute to her the belief she avows.[41]

However, there are a variety of ways in which a person's deeds can be at odds with his words, and a corresponding variety of situations in which an avowal deserves the retort, "You don't really believe that." I think that in most, if not all of them, there is scope for saying that the avowal was not sincere or transparent. We may safely ignore here avowals that involve slips of the tongue or malapropisms, as well as anaphoric avowals ("I believe that, too") that are based on mishearing or misunderstanding what was just said. In such cases the speaker does not have the belief he says he has, but this is simply because he didn't say what he meant to say. A much more significant class of cases includes avowals that are made when the speaker is in the grip of an emotion that colors his judgment. Consider the following case.[42] A young man has recently moved to Detroit, and in a conversation with his father the son says, "I've come to think that Detroit is a pretty horrible place—ugly and harsh. I will probably move to New York this year." Then in a second conversation not long afterward, the son says, "There is a constant sense of challenge and 'edge' in Detroit. I love the place." Of course, the young man may simply have changed his mind about the city. But a more interesting possibility is that the father knew all along that the first avowal did not really reflect what his son believed. And this might have been the case even though there was no reason for thinking that the avowal was insincere.

What is crucial here is that the young man's emotional state was coloring his judgment about Detroit. And this implies, I think, that there was at least a partial failure of transparency in his initial avowal of belief. While the son had one eye on the facts about the city (Detroit, after all, isn't Paris), his avowal reflected in part feelings of disappointment, temporary loneliness, or the disorientation that accompanies change. It would be wrong to say that the avowal was insincere, for it represented a sincere attempt to fit words to this complex of affective and cognitive attitudes.[43] But his emotional state led him to exaggerate certain features of the city and to discount others. The father's judgment that his son didn't really think that Detroit was all that horrible presumably reflected his awareness that were his son to focus more clearly on the facts about his new home, and treat the question about what he believed as fully transparent to these facts, he would render a more favorable judgment on it.[44]

Moving on to more seriously deficient avowals of belief, we enter the area of self-deception. I want to suggest that many instances of this phenomenon involve avowals that are, in a certain sense, insincere. There is often some form of *pretense* at work in such cases, and here it must be born in mind that there are subtler forms of pretense than simple lying. Some of them are so subtle that the pretender is unconscious of them, giving rise to failures of self-knowledge.

Consider, for example, the following mild case of self-deception. Sam is a young professor of philosophy attending a reception for the new dean of his college, who is from the comparative literature department. A small group of faculty from various departments are extolling the virtues of Derrida's work, and bemoaning the fact that analytic philosophers do not appreciate him. When the dean asks for Sam's opinion, he says, "I think analytic philosophers have been a bit too hard on Derrida," whereupon Sam's friend and colleague Jane rolls her eyes and mutters under her breath, "You don't really believe that!" When Jane taunts Sam about his remark afterward, he initially tries to defend himself, reiterating that he does think that many philosophers have been a *bit* too hard on Derrida's work, but he eventually admits that Jane is right. Sam was merely going along with the crowd, trying to ingratiate himself to the new dean, and he feels duly ashamed.

How was Jane able to see that Sam didn't really believe what he claimed to believe? Presumably she has never before heard Sam say anything sympathetic toward Derrida, and has ample evidence for thinking that Sam thinks that most French philosophy is incoherent and valueless. Of course, Sam might recently have changed his mind on this subject, but Jane rightly judged that Sam had no intention of living up to the claim he endorsed. This was made easier by the availability of an alternative explanation for his avowal, namely, that it was an attempt to please his interlocutors. Sam wasn't lying; his pretense was neither conscious nor calculated. But it was pretense nonetheless, given that the best explanation for his remark would not cite the fact that he thought Derrida's work had some value, but rather the fact that he wanted to *seem* to be someone who thought Derrida's work had some value. Perhaps Sam has a habit of agreeing with others in order to avoid conflict or secure advantages for himself. This is a character flaw, which explains his occasional lapses of self-knowledge better than the postulation of some cognitive deficit.

For another example, suppose that Jake is a congressman who professes to believe that discrimination on the basis of race is impermissible, but routinely passes over qualified black people who apply for jobs in his office, and retains his membership in an all-white country club. How we should characterize Jake's beliefs depends largely on how he responds when the inconsistency between what he says and what he does is called to his attention. If he acknowledges the inconsistency and vows to amend his ways, then perhaps he really does believe what he claims to believe. His behavior may simultaneously warrant saying that he still also believes that in certain circumstances discrimination on the basis of race is permissible. That is, he may have contradictory beliefs, which is a failure of self-knowledge of a sort, though not of first-person authority.[45] On the other hand, suppose that Jake tries to brush aside or downplay the significance of his actions when

their inconsistency with his avowals is pointed out to him. This would probably indicate that he doesn't really believe what he says, but at the same time it would force us to look for an alternative explanation for his saying it. Perhaps he *wants* to be the sort of person who treats black and white people equally. This is a virtuous ambition, but it is not the right sort of basis on which to avow a belief. In doing so he is pretending to have arrived at where he wants to go; his mind is on an image, not on reality, so his avowal is not sincere, or at least, not both sincere and transparent.[46]

For still another possibility, suppose Jake tends to pretend to colorblindness most frequently and forcefully when he is arguing against affirmative action programs for minorities. We may then have reason to question not only whether he believes what he says, but even whether he sincerely wants to believe it. Perhaps he merely wants to *seem* to be the sort of person who believes that justice should be colorblind. Yet the context might be such that it would be incorrect to call his professions of belief plain lies. What is probably at work here is *hypocrisy,* of which Elizabeth Anscombe remarks, "It is characteristic of this sort of wanting-to-seem that it carries with it an implicit demand for respect for an atmosphere evoked by the pretender, which surrounds not the reality, but the idea of such things as being principled, or cultured, or saintly, or rich, or important."[47] The role played by such factors in providing the context for the hypocrite's avowals clearly marks them as judgments that are not sincere and transparent. This is what explains the failures of self-knowledge these judgments manifest.

Human beings are naturally imitative creatures. Young people frequently go through a period in which they try on various roles, parroting the words of a favorite teacher or other admired figure, while not fully understanding their implications, and hence not being in a position to live up to the commitments they express. They begin to attain self-knowledge when they learn to think for themselves—about matters in general, not just about matters having directly to do with themselves—and this involves the development of character as much as cognitive skills. Should they instead take into adult life the habit of mimicking those around them, they may lose the capacity to make sincere and transparent judgments that is the basis of first-person authority.

Along similar lines, it may be noted that one of the pathologies engendered by social oppression is the destruction of the self-knowledge of the oppressed, as a result of their being forced to play the roles allocated to them by the oppressors in order to survive. Of the subjection of women under patriarchy, Adrienne Rich writes, "the lie of the 'happy marriage,' of domesticity—we have been complicit, have acted out the fiction of a well-lived life. . . . There is a danger run by all powerless people: that we forget we are lying, or that lying becomes a weapon we carry over into our relationships with

people who do not have power over us."[48] The restoration of the self-knowledge of the oppressed is ultimately to be achieved through the creation of the conditions in which they can express themselves without fear of reprisal. This is a problem for politics, not cognitive psychology.

Attending to the respects in which self-knowledge is a practical rather than a cognitive achievement affords the best explanation of why self-knowledge is regarded as an ethical as well as an epistemic virtue. It is often difficult to resist the pressures operating that demand assent to prevailing pieties; it can require considerable courage to look at the world with one's eyes and make the sincere, transparent avowals that manifest self-knowledge. And it's harder for the poor and powerless to know themselves for the same reasons that it's harder for them to live well.

VII

Recently there has been a resurgence of interest in first-person authority, spurred largely by the widespread acceptance of anti-individualism about mental content, the compatibility of which with first-person authority has seemed doubtful to some.[49] If the contents of a person's mental states are determined in part by environmental conditions, then mustn't his mental self-ascriptions, if they are to count as knowledge, be founded in part on reasons or evidence for thinking that the relevant environmental conditions obtain? If so, then since such reasons or evidence could only be obtained via observation, in a broad sense, it is hard to see how avowals could be authoritative.

Significantly, anti-individualism has never been seen as a threat to an individual's ability to *have* first-order attitudes, or to make first-order judgments, with determinate contents. The worry is that one's second-order judgments *about* those first-order attitudes cannot manifest direct and authoritative knowledge. This has led to one line of anti-individualist thought on first-person authority, which seeks to articulate constitutive connections between the contents of a person's first- and second-order attitudes, in support of the claim that of necessity, at least typically, the latter accurately reflect the former and can for this reason be regarded as constituting knowledge.[50]

The reasoning that leads to doubts concerning the compatibility of anti-individualism and first-person authority is questionable on more general grounds. It implicitly requires that the knowledge a person has of his non-individualistically determined thought contents should itself be explicable individualistically, in terms of certain kinds of reasons or evidence available to the subject. But the point of at least one strand of anti-individualistic thought is that the norms governing human thought may transcend an individual's ability to articulate them. This surely ought to apply to the norms governing the concepts of knowledge and warrant as well as those of belief, desire, and intention. Anti-individualism about mental content thus gives rise to a need for an anti-individualistic account of first-person authority. Moreover, and even setting aside anti-individualism, it is intuitively implausible that the warrant a person possesses for her mentalistic self-attributions typically consists in reasons or evidence that support them. The knowledge manifest in avowals is too direct for that.

The expressivist account of first-person authority as it applies to avowals of belief presented in this paper seeks to incorporate both these ideas. It is genuinely anti-individualistic in holding that the authority of such avowals is only intelligible when they are seen in the context of the social practices in which they have their home. Reflection on these practices shows that paradigmatic avowals are qualified assertions, which provides for precisely the kind of constitutive connection between first- and second-order judgment alluded to in the first line of anti-individualistic response mentioned above. Indeed, we have no need to distinguish first- and second-order beliefs when a sincere, transparent avowal is in question; the judgment, "I believe it is raining," typically expresses directly the subject's belief that it is raining, rather than a higher-order judgment about that belief. But it does so via a sentence that semantically ascribes that very belief to the subject, so that avowals of this form are true, provided they are sincere and transparent. While not all self-ascriptions of belief have these features, those that do not are not as fundamental, nor as authoritative, as those that do. There is no danger that expressivism about avowals lapses into the doctrine that "saying makes it so," for a person's avowals are still accountable to the available third-person evidence for and against the attribution, which may on occasion render a verdict about the subject's mind that is at variance with his own. But most, if not all, of the kinds of error to which self-ascriptions of belief are prone can be seen as failures of sincerity or transparency, in accordance with the pre-theoretical intuition that mistaken self-ascriptions are typically due to some criticizable failure on the part of the subject. A person's avowals are authoritative to the extent they reflect the autonomy of a rational agent to shape his thinking in accordance with the norms of reason.

NOTES

This paper is a descendant of a paper presented at UC Santa Barbara and UCLA in February 1993. For helpful comments and criticism, I am grateful to Robert Adams, Tony Anderson, Johannes Brandl, Tony Brueckner, Tyler Burge, Francis Dauer, Kit Fine, Matthew Hanser, Chris Hill, Andrew Hsu, Robin Jeshion, David Kaplan, Gene Mason, Joseph Owens, Sandra Peterson,

Michael Root, Nathan Salmon, Voula Tsouna, John Wallace, and Richard Warzecha.

1. For some recent work on our epistemic right to accept the testimony of others, see C. A. J. Coady, *Testimony: A Philosophical Study* (Oxford: Oxford University Press, 1992), and Tyler Burge, "Content Preservation," *Philosophical Review* 102 (1993): 457–88.
2. I thus disagree with Gilbert Ryle, who saw first-person authority as simply a special case of the fact that persons are generally reliable interlocutors, especially on subjects with which they are very familiar, which would of course include themselves. See Ryle's *The Concept of Mind* (Harmondsworth: Penguin, 1949), chap. 6.
3. Tyler Burge has suggested that avowals are not subject to "brute error," which he characterizes as error that does "not result from any sort of carelessness, malfunction, or irrationality on [the subject's] part" ("Individualism and Self-Knowledge," *Journal of Philosophy* 85 [1988]: 649–63; 657). I am in general agreement with this, though in discussing failures of self-knowledge, I tend to emphasize certain culpable kinds of insincerity. See section VI below.
4. Wittgenstein, *Philosophical Investigations* (Oxford: Blackwell, 1953), sec. 244.
5. J. J. C. Smart, "Sensations and Brain Processes," *Philosophical Review* 68 (1959): 141–56; 144. It is not clear whether Smart endorses this way of viewing avowals of love. His primary aim is to deny the corresponding thesis about avowals of sensations.
6. This is simply an adaptation of the reasoning used in Peter Geach's "Assertion," *Philosophical Review* 65 (1965): 449–65, to cast doubt on ethical emotivism.
7. In "Sensations and Brain Processes," Smart introduces the expressivist account of avowals of love mentioned above by contrasting it with the idea that "'I love you' [is] normally a report that I love someone" (144).
8. It is not clear to me that avowals of sensations cannot be construed as reports of inner states.
9. I do not think Wittgenstein made this mistake, though the matter is complex. Wittgenstein denied that "I am in pain" expresses knowledge, but he seems to have been led to this view because he thought that one could speak of knowledge only where there existed the possibility of error, which he thought absent in this case. I do not know of any place in his writings where he suggests that one doesn't typically know what one believes or intends, and I do not think he intended to deny that avowals are assertions to which truth and falsity apply. Moreover, his flirtations with deflationism about truth would seem to preclude his adoption of the metaphysical standpoint from which one could divide apparently similar regions of discourse into those that do and those that do not genuinely purport to state facts. On the other hand, the idea that meaning is, or is determined by, use is not congenial to distinguishing between semantics and pragmatics as I have here. Such a distinction is adequately motivated, I think, by the need to find sufficient common structure in utterances that serve different communicative purposes to ensure the validity of arguments that incorporate them. As for Ryle, his remarks about "unstudied utterances" sometimes seem aimed at denying that avowals of belief are assertions *about the speaker* (to this extent his view may be compared to the view of J. O. Urmson discussed below in n. 14). But I do not think Ryle ever claimed that avowals do not manifest self-*knowledge;* he was primarily concerned to deny that it was a distinctive kind of knowledge.
10. Thus, the cognitivist-expressivist version of Wittgenstein's thesis about pain would yield the thesis that sincere utterances of "I am in pain" are true. I take no stand on this thesis. As for the corresponding thesis about love, I think it is very problematic. See n. 45 below.
11. *Philosophical Investigations,* sec. 647. The passage is criticized by Elizabeth Anscombe in *Intention* (Oxford: Blackwell, 1957), 5: "One might as well call a car's stalling the *expression* of its being about to stop."

12. All of these arguments that knowledge is required for an assertion to be warranted, as well as a number of others, may be found in Timothy Williamson, "Knowing and Asserting," *Philosophical Review* 105 (1996): 489–523.
13. It will occasionally be convenient in what follows to employ the phrase 'assertion that *p*' broadly, to cover both the unqualified assertion "*p*" and the qualified assertion "I believe that *p*." J. O. Urmson, in "Parenthetical Verbs," *Mind* 61 (1952): 480–96, drew attention to the fact that "I believe that *p*" is ordinarily used to assert that *p* in a qualified way, but he thought this showed that verbs such as 'believe' are not "psychological descriptions." His mistake here is once again that of drawing a semantic conclusion from a pragmatic observation. Urmson seems to be committed to the implausible view that an utterance of (i) is true if and only if the library is two blocks down on the right. This of course makes it impossible to account properly for the entailments of (i). In my view the speaker who asserts (i) does make a qualified assertion to the effect that (iii), but he also gives a psychological self-description, asserting that he has a certain belief. This is the proposition semantically expressed by his utterance, and it is true if and only if the speaker believes that the library is two blocks down on the right. Arthur Collins, in *The Nature of Mental Things* (Notre Dame: University of Notre Dame Press, 1987), also notes that "I believe that *p*" commits the speaker, at least provisionally, to the truth of *p.* But he argues on this basis that the semantics of this sentence is not what it appears to be. I think this latter claim is mistaken. Collins is criticized on this score by Richard Moran, "Arthur Collins's "The Nature of Mental Things," *Philosophy and Phenomenological Research* 87 (1994): 917–20.
14. Of course, if one is asked for the location of the library, it would ordinarily be odd to assert (ii) and then go on to give a different answer of one's own. But that is only because asserting (ii) in this context would ordinarily be seen as deferring to the friend on the basis of the fact that one has no helpful information of one's own. Moreover, it isn't always odd to report what the friend thinks and then give a different answer of one's own. Consider, for example, "My friend thinks it's two blocks down on the right, but since he habitually underestimates distances, it's probably three or four blocks down on the right." In contrast, the result of replacing the words 'my friend' and 'he' in this assertion with the word 'I' would be bizarre indeed.
15. The significance of Moore's paradox for an understanding of the authority of avowals was first noticed by Wittgenstein. *Philosophical Investigations*, pt. 2, sec. 10, and his remarks on the paradox were the stimulus for the main ideas of this paper. Moore's paradox has been put to use in rather different treatments of self-knowledge in Sydney Shoemaker, "On Knowing One's Own Mind," *Philosophical Perspectives* 2 (1988): 183–209, and Andre Gallois, *The World Without, The Mind Within: An Essay on First-Person Authority* (Cambridge: Cambridge University Press, 1996). Although she is primarily interested in the paradox itself, and only secondarily in first-person authority, the discussion of the authority of avowals in Jane Heal, "Moore's Paradox: A Wittgensteinian Approach," *Mind* 103 (1994): 5–24, is quite similar in spirit to the present account. But I disagree with Heal's claim that the authority of avowals is "not epistemic."
16. That the speaker who asserts (iii) may be mistaken or lying does not alter the fact that the *role* of this assertion is to express knowledge. If I mistakenly or deceitfully assert (iii), I am misusing the assertion. On this point, see Williamson, "Knowing and Asserting."
17. Possible Gettier cases will be mentioned in the next section.
18. Recall that I am using the term *assertion* widely, in such a way that the assertion "I believe that *p*" counts also as a qualified assertion of "*p.*"
19. I appealed only to the left-to-right half of this biconditional in the last section. It is the right-to-left half I think false.
20. But I did not *lie* to her, because I did not tell her anything I believed to be false. Other examples of assertions intended to deceive that fall short of lying include the strategic assertion of half-truths or otherwise misleading facts. Still other kinds of insincerity will figure importantly in section VI.
21. Of course, given that what I foresee to be the likely consequence of my advice is that

someone who has a low opinion of me will act in such a way as to do herself a minor harm and me a benefit, there are ample opportunities for self-deception here. But only a cynic would think that the interpretation I gave of my own action in this case could not possibly be taken at face value.

22. The possibility of calculated, insincere assertions that coincidentally accord with what the speaker actually thinks brings with it the possibility of a kind of Gettier case in which an individual forms a warranted true belief that is not knowledge about what another person thinks. Suppose once again that Jill asks me which route to take to L.A., I think that 101 would be best, and I think that she thinks I am a fool. But now suppose that this last belief of mine is false; in fact Jill has asked for my advice because she believes I know about such things. I respond as in the first example above, cleverly disguising the calculated nature of my response, and Jill goes away thinking that I think that 101 is the best route. Her belief about what I believe is true, and warranted as well, since she had no reason to think that the normal conditions for reliance on first-person authority did not obtain in this case. But because of my insincerity, one of these conditions did not in fact obtain, so her belief is not knowledge. Jill's tacit but mistaken acceptance of my sincerity plays a role here analogous to the false premises in inferences in the standard Gettier examples. By the same token, even if I say flatly, "101 is the best route," and I know this to be true, if this is a calculated attempt to deceive her it does not transmit my knowledge to her, though she will acquire a warranted true belief that 101 is the best route.
23. The etymology of the word *sincerity* also suggests this way of looking at the notion. It comes from the Latin *sin cera,* "without wax"; that is, without polish or adornment, or as we might say in a contemporary colloquialism, without "spin." For this and other material on the history of the concept of sincerity, see Lionel Trilling, *Sincerity and Authenticity* (Cambridge, Mass.: Harvard University Press, 1972).
24. Gareth Evans, *The Varieties of Reference* (Oxford: Oxford University Press, 1982), 225.
25. The term 'transparency' is due to Richard Moran, who has emphasized the importance of the phenomenon. In his "Making Up Your Mind: Self-Interpretation and Self-Constitution," *Ratio* 1 (1988): 135–51, Moran points out that to say that one question is transparent to another is not equivalent to saying that the one reduces to the other. One question reduces to another when the answers to both are determined by the same fact. But in general, that I believe (or disbelieve) that p is not true in virtue of the same fact that makes it true (or false) that p. See also Moran's "Interpretation Theory and the First Person," *Philosophical Quarterly* 44 (1994): 154–73.
26. In *The World Without: The Mind Within,* Andre Gallois seems to hold that transparency is lacking when memory is relied on in an avowal (see esp. 114–15). My criticism of this view in the next paragraph adapts to the present topic a conception of the role of memory in deductive reasoning that is set forth in Burge's "Content Preservation."
27. Memory does play a justificatory role with respect to self-attributions of *past* beliefs, such as "I thought last week that the dissertation was ready." For this reason, I do not think that such judgments are as authoritative as avowals of present beliefs. The role of memory in avowing standing beliefs is similar in certain respects to the role of testimony in the production of beliefs. When I form the belief that it is raining on the basis of someone's telling me so, she does not tell me that I think it is raining, she tells me that it is raining (perhaps adding the qualifier, "I think"), and I endorse this content (though I can also refrain from endorsing it). Similarly, memory does not "tell" me that I think the student's dissertation is ready; it is the voice of my former self "telling" me (perhaps with qualification) that the work is ready, enabling me to re-endorse this content.
28. I believe I have here reached, albeit by a rather different route, a conclusion similar to Shoemaker, who argues in "On Knowing One's Own Mind" that self-knowledge is "supervenient" on normal conceptual capacity and rationality. It should be noted that I am not suggesting either that (a) one's warrant for judging that one thinks that p consists of whatever warrants one in thinking that p, or that (b) one infers that one believes that p from the fact that one has judged that p. The idea that there is an inference involved here strikes me as phenomenologically inaccurate; nor do I see how such an inference could be justified. The following weakened version of (a) also strikes me as dubious: (a′): one is warranted in judging that one believes that p only if one is warranted in believing that p. If one unwarrantedly judges that p out of, say, simple gullibility, the judgment might still represent a sincere attempt at stating a truth, so the corresponding self-attribution would be warranted. On the other hand, if one professes to believe that p while cavalierly disregarding substantial evidence that not-p, we should probably begin to wonder whether the subject is treating the question concerning what he believes as transparent to the question whether p, and therewith the presumption of first-person authority would indeed be undermined.
29. That Evans's observation must be qualified in this way is persuasively argued in Richard Moran's "Making Up Your Mind: Self-Interpretation and Self-Constitution."
30. It might be suggested that genuinely introspective self-ascriptions of belief of the kind considered here are not really *avowals* at all, inasmuch as they lack the requisite spontaneity. This may well be correct, but I would resist the correlative suggestion that the term 'transparent avowal' is redundant. In section VI we will find reason to allow for non-transparent avowals in connection with certain kinds of self-deception.
31. Ryle, *The Concept of Mind,* chap. 6.
32. I venture to suggest that Ryle made the mistake of failing to distinguish between the *phenomenon* of first-person authority, the main features of which were canvassed at the beginning of this paper, and the Cartesian *theory* introduced to explain the phenomenon. As a result, in his zeal to throw out the bathwater of "privileged access" Ryle threw out the baby of first-person authority along with it.
33. In fact, the Moore-like utterance of the client in psychotherapy mentioned two paragraphs back teeters on the brink of unintelligibility. The realization it is trying to express would be better put in the form, "I have believed all this time that my father was cruel to me, but he wasn't," where the past tense in the first conjunct indicates that because he now sees that his father wasn't cruel to him, the speaker no longer believes that he was.
34. Psychotherapists are familiar with the spurious "insights" that clients often come up with after prolonged introspection. They are often the product of a self-deceptive fear that if the therapy continues it will unearth still more unpleasant truths.
35. Stuart Hampshire criticizes the "ideal of sincerity as mere naturalness," and argues that some form of introspection ("watching oneself") is a necessary condition of sincerity. Given the assumption that "there are less than conscious thoughts, . . . which may be in conflict with the thoughts . . . which are fully conscious," Hampshire writes, "self-watching is always necessary as a precaution against an unrecognized conflict or confusion of thought" (see Hampshire's "Sincerity and Single-Mindedness," in *Freedom of Mind* [Princeton: Princeton University Press, 1971]). But it does not seem to me necessary for a person to adopt the critical, introspective stance unless and until one has reason for thinking that some unconscious conflict is operative in a particular instance. Compare: there are papier mache facsimiles of barns, but in ordinary circumstances I can justifiably claim that an object is a barn without having inspected it closely to rule out the possibility that it is a papier mache fake.
36. Shoemaker has dubbed this the "broad perceptual model," and argues at length against it in "Self-Knowledge and 'Inner Sense,'" *Philosophy and Phenomenological Research* 54 (1994): 249–90. For other arguments against the perceptual model, see Donald Davidson, "Knowing One's Own Mind," *Proceedings of the American Philosophical Association* 60 (1986): 441–58, and Tyler Burge, "Our Entitlement to Self-Knowledge," *Proceedings of the Aristotelian Society* 96 (1996): 91–116.
37. See the discussion of the authority of observation reports in Sellars, "Empiricism and the Philosophy of Mind," in *Science, Perception, and Reality* (New York: Humanities Press, 1963), sec. 4.
38. The perceptual model appears powerless to explain a crucial difference between beliefs and another important type of content-bearing state, namely, perceptual states. As Wittgenstein remarked (*Philosophical Investigations*, pt. 2, sec. 10), "One can doubt

one's perception, but not one's belief." That is, while I may judge, "It looks to me as though the top line is longer, but I doubt that it is," it is not intelligible to judge, "I believe that the top line is longer, but I doubt that it is." It is possible to adopt an essentially third-person stance toward one's own perceptual systems, treating them as one treats the testimony of another person, registering the contents of their representations without endorsing them. Thus, it may be appropriate to account for one's knowledge of one's own perceptual states in terms of a kind of observation. But such a third-person stance is not possible with respect to one's own beliefs. This is readily explainable on the expressivist account of belief avowals, according to which they involve a commitment to the truth of the content proposition.

39. Boghossian, "Content and Self-Knowledge," *Philosophical Topics* 1 (1989): 5–26.
40. The notion of "cognitively insubstantial" self-knowledge is due to Crispin Wright, "Wittgenstein's Rule-Following Considerations and the Central Project of Theoretical Linguistics," in *Reflections on Chomsky,* ed. Alexander George (Oxford: Blackwell, 1989), and elsewhere. Boghossian's criticism is directed primarily at Burge's observations in "Individualism and Self-Knowledge" on the self-verifying character of avowals of occurrent thoughts. Burge's work on this topic was an important stimulus for this paper.
41. In an interesting paper, Bernard Kobes argues that avowals of belief should be seen as having the direction of fit characteristic of expressions of intention. A sincere avowal of belief is *made true* by the speaker's subsequent judgments and actions. See Kobes, "Mental Content and Hot Self-Consciousness," *Philosophical Topics* 24 (1996): 71–99.
42. My thanks to Chris Hill for suggesting this case.
43. It is unfortunate that the verb 'feel' is increasingly usurping the place of 'think' or 'believe' in ordinary language. This case illustrates the need to keep these notions distinct. For notice that if the young man's initial avowal had been, "I feel that . . ." rather than "I think that . . . ," it would have been not merely sincere but true: that is what he was feeling at that time. His only error was in presenting his judgment as solely the product of thought with no admixture of feeling.
44. Cases such as these should lead us to conclude that avowals of attitudes such as love—which by its very nature consists of a complex of affective, cognitive, and practical components—are much less authoritative than avowals of belief. I argued in the first section that seeing utterances of "I love you" as expressive of love is no barrier to treating them as genuine assertions. On the other hand, I do not think that any principle analogous to (E2') can be stated for avowals of love. Sincerity is clearly insufficient for truth in this case, and I do not think that the notion of transparency can be applied to avowals of love. There does not seem to be any more basic or ground-level question to which the question, "Do I love Susan?" ought, ideally, to be transparent. There is inevitably scope for introspective self-interpretation when one entertains this question.
45. The undetected presence of contradictory beliefs is a failure of what Joseph Owens and I call "introspective knowledge of comparative content" in our "Externalism, Self-Knowledge and Skepticism," *Philosophical Review* 103 (1994): 107–37. Note that I have not in this paper made any claims regarding the authority of *disavowals* of belief. I am inclined to think that disavowals are significantly less authoritative than avowals.
46. I am not sure that it is possible or necessary to isolate the failure here as consisting simply in insincerity or non-transparency. What I think is clear is that it is not the case that both of the conditions that (E2') requires for warrant are satisfied.
47. Elizabeth Anscombe, "Pretending," in *Metaphysics and Philosophy of Mind: Collected Papers,* vol. 2 (Minneapolis: University of Minnesota Press, 1981). I follow her in reserving the term *hypocrisy* for a certain type of "non-plain" pretense; that is, pretense in which the pretender does not "unreflectively" know he is pretending. Thus, hypocrisy typically involves a lapse of self-knowledge, and Captain Renault was not, as is sometimes said, being hypocritical when he closed Rick's Cafe in the film *Casablanca,* since he knew perfectly well he wasn't really shocked that gambling was going on there. Renault's performance was rather a plain pretense of hypocrisy—doubly embedded pretending, as it were—which, as Anscombe observes, "is one of the popular senses of *cynicism,* . . . and is found, e.g., among the clearer-headed politicians."
48. Adrienne Rich, "Women and Honor: Some Notes on Lying," *Adrienne Rich's Poetry and Prose,* ed. B. C. Gelpi and A. Gelpi (New York: W. W. Norton, 1993). She is here using the word 'lie' in the extended sense it occasionally has in ordinary language, which corresponds to my use of the word 'pretense.' One can "lie" in this sense, without being conscious of it, and even while believing what one says (Rich writes, "It has been difficult, too, to know the lies of our complicity from the lies we believed."). I take it that Rich would maintain that there are situations in which a woman says, for example, "I think my husband has been good to me," and this is true (both in the sense that it is what she thinks and in the sense that he has been good to her), but where the best explanation of her saying this would mention the social pressure upon her to do so. Hence the applicability of the notion of pretense ("lie"): her avowal is not sincere and transparent. Even though it is true, it does not manifest self-knowledge. Obviously, Rich's remarks could be applied, *mutatis mutandis,* to various groups of individuals who have suffered from the effects of systematic racial or ethnic discrimination.
49. The seminal works on anti-individualism are Hilary Putnam, "The Meaning of 'Meaning,'" in *Mind, Language, and Reality: Philosophical Papers,* vol. 2 (Cambridge: Cambridge University Press, 1975), and Tyler Burge, "Individualism and the Mental," *Midwest Studies in Philosophy* 5 (1979): 73–121. For doubts about the compatibility of anti-individualism and first-person authority, see, for example, Boghossian, "Content and Self-Knowledge," and Anthony Brueckner, "Knowledge of Content and Knowledge of the World," *Philosophical Review* 104 (1994): 327–44.
50. For this type of strategy, see Burge, "Individualism and Self-Knowledge," and "Our Entitlement to Self-Knowledge," and, in a rather different vein, Shoemaker, "On Knowing One's Own Mind." For a general discussion of this "level-linking" strategy, see Christopher Peacocke, "Entitlement, Self-Knowledge, and Conceptual Redeployment," *Proceedings of the Aristotelian Society* 96 (1996): 117–58.