From Transparency to Expressivism[1]

David H. Finkelstein

1. At least since John Locke defined "*REFLECTION*", as the "*Perception of the Operations of our own Minds* within us",[2] one influential approach toward understanding the knowledge we have of, and the authority with which we speak about, our own psychological states and events has been what I call "detectivism". A detectivist is someone who would claim that we know our own mental goings-on thanks to a process by which we perceive or, anyway, detect them and thereby render them conscious. So characterized, "detectivism" names a broad family of positions. While some detectivists claim that self-knowledge comes via an "inner sense" that provides each of us with infallible epistemic access to a peculiarly private, perhaps immaterial, mental realm, others (of a more naturalistic bent) posit a perceptual process that is epistemically on all fours with seeing and hearing, except that it happens to be directed inward, toward the mind/brain rather than outward toward the external world.[3]

Detectivism as such is controversial; many philosophers and psychologists are committed to one or another version of it, while many others reject it. In the present essay, I won't be engaging directly with arguments either for or against detectivism. Instead I want to consider a way of answering the following question: *If* we reject detecti-

1 Versions of this paper were presented at the University of Chicago's Wittgenstein Workshop in March 2010; at a conference called Self and Others in Wittgenstein and Contemporary Analytic Philosophy at the University of Southampton in March 2010; at the Inter-University Workshop on *Expression and the Inner*, held in Oviedo, Spain, in April 2010; at a conference called Aspects of Self-Knowledge at LMU München in December 2010; at Auburn University's Department of Philosophy in February 2011; and at Johns Hopkins University's Department of Philosophy in September 2011. For especially helpful comments, questions, or conversation, I'm grateful to Anita Avramides, Stina Bäckström, Matthew Boyle, Jason Bridges, James Conant, Adrian Haddock, Arata Hamawaki, Jane Heal, Irad Kimhi, Thomas Lockhart, Robert Pippin, William Small, and Barry Stroud.

2 Locke (1689/1975), II, I, 4.

3 For a selective history of detectivism, see ch. 1 of Finkelstein (2003).

vism, how are we supposed to make sense of the fact that people are able to accurately and easily self-ascribe beliefs, desires, fears, hopes, and the like? How can we give up detectivism without this capacity just coming to seem miraculous? The answer that I'll be discussing in the bulk of what follows is developed in the work of Gareth Evans, Sydney Shoemaker, and Richard Moran. According to (what we might call) the transparency approach toward understanding self-knowledge and first-person authority, we are able to speak about our own states of mind by addressing questions about the world outside us. Where this is possible, we don't need to rely on any sort of inward-directed perception or detection in order to say what we ourselves want, fear, hope, or believe.

This is not the first time that I have written about the transparency approach. In a postscript to Finkelstein (2003), I criticized it, focusing especially on the innovative version that Moran had defended in his 2001 book, *Authority and Estrangement*. In a paper published in 2009, Matthew Boyle offered (what struck me as) a sensible reply to my criticism of Moran. When I read Boyle's paper, I decided that I needed to consider again the question of precisely what, if anything, I take to be wrong with Moran's position—along with the transparency approach, more generally. The present essay is a result of that reconsideration.

2. Evans writes:

> [I]n making a self-ascription of belief, one's eyes are, so to speak, or occasionally literally, directed outward—upon the world. If someone asks me 'Do you think there is going to be a third world war?', I must attend, in answering him, to precisely the same outward phenomena as I would attend to if I were answering the question 'Will there be a third world war?' I get myself into a position to answer the question whether I believe that *p* by putting into operation whatever procedures I have for answering the question whether *p*.[4]

We can capsulize Evans's point in the following claim, which I'll refer to in what follows as "TC" (for "transparency claim"):

> TC: The question of whether I believe that *p* is, for me, transparent to[5] the question of whether *p* is true. This is to say: I can answer the former question by answering the latter.

4 Evans (1982), 225.

5 In this use of the phrase "transparent to", I'm following Moran (2001).

Suppose that TC strikes us not only as true, but as showing that we need not posit an inner sense in order to make sense of self-ascriptions of belief. Suppose, moreover, that we would like to extend the thought expressed in TC beyond belief (as it were) and to thereby explain the authority with which we speak about our own desires, fears, hopes, and other attitudes. It's far from obvious how such an extension might be effected. Shoemaker tries and, it seems to me, fails with various proposals. For example, he writes:

> The rational agent who wants X and has normal mastery of language will, ceteris paribus, respond affirmatively to the question "Shall I give you X?" And given her mastery of the concept of desire, she will respond affirmatively to the question "Do you want X?" if she will respond affirmatively to the question "Shall I give you X?" So she will, unless she has devious motives, give correct answers to questions about what she wants.[6]

Here, Shoemaker is suggesting, in effect, that if you were to ask me a question of the form "Do you want X?", I could answer by acting as if you had asked me a corresponding question of the form "Shall I give you X?". But this suggestion seems blatantly unsatisfactory. I might want anything—a sandwich, a summer house, a fulfilling career, or an ability to yodel—without wanting you (or perhaps anyone) to give it to me.

As I read Moran, he offers another way to extend the kind of explanation of first-person authority that is suggested by TC (to extend it, that is, to attitudes other than belief). And it seems to me that what he provides along these lines is, *prima facie*, a good deal more appealing than what Shoemaker offers. Moran manages this by shifting his starting point from TC to what I'll call TC*:

> TC*: The question of whether I believe that *p* is, for me, transparent to the question of what I ought rationally to believe—i.e., to the question of whether the reasons require me to believe that *p*. I can answer the former question by answering the latter.

The difference between TC and TC* lies in TC*'s appeal to reasons.[7] TC* says, in effect, that if I am asked whether I believe it's about to

6 Shoemaker (1994), 283.

7 Moran (2003), 405, writes: "[T]here's something a little misleading in Evans's original formulation. He speaks of one's eyes 'directed outward, upon the world', and contrasts this outward gaze with the discredited inward one. But it is rarely by simply *looking* at something that we answer such world-directed questions for ourselves, and when looking is involved it is because in this con-

rain, I can address this question by answering the question "Do the reasons require me to believe that it's about to rain?". This shift in starting points has the effect of making TC* straightforwardly extendable to other attitudes. According to Moran, the question of, e.g., whether I want X is transparent to the question "Do the reasons require me to want X?"; I can say what I want by consideration of what I ought to want in light of the appropriate reasons. Or, rather, I can do this on the condition that what I in fact want is, after all, in accord with my assessment of what I ought to want. (Moran calls this the "Transparency Condition".) When, and only when, this condition is in place, I can say what I want by considering what I ought to want, what I fear by considering what I ought to fear, what I intend by considering what I ought to intend, etc.

Moran uses the word "avowal" only where a person's answer to a question about some attitude of hers is dictated by her assessment of what is rationally required of her. Avowal, in this sense, is for Moran the essence of first-person awareness. He understands "the ability to avow one's belief as the fundamental form of self-knowledge".[8] He holds that a person has "genuine first-person awareness"[9] only when she can avow (in his sense) her attitude—only when she can speak about it by addressing the outward-directed question of what the appropriate reasons require of her.[10] In the postscript to my book, I argued

text we are counting what we see to provide some reason for or against some possible belief. The more general and central truth in this context is that I answer the external question about the weather or the possibility of war by putting myself in a position to confront and assess the reasons relevant to the truth about the weather or the possibility of war."

8 Moran (2001), 150.

9 Moran (2001), 107.

10 Boyle (following Moran) sometime speaks as if Moran's central idea is that I know my beliefs, intentions, fears, etc., by making up my mind, or by deliberating. Thus Boyle writes: "What Moran finds striking about our knowledge of our own attitudes is this: we often seem to be able to know whether we hold them *by deliberating about the topics they concern*" (Boyle (2009), 136, his emphasis). This way of putting the point is, I think, liable to mislead. Here's what's misleading: I often say what my attitude is—e.g., what I believe—when I'm *not* making my mind, when my mind is already made up, and I don't now need to deliberate. (If I were asked who I thought the president of the United States was, I could answer without needing to deliberate or make up my mind.) One might conclude that Moran cannot accommodate self-knowledge in such cases. But I don't think this is a compelling objection to Moran's view, rightly understood. For his view could be put this way: regardless of whether I'm making up my mind or merely maintaining it (rationally), I answer the question of what my attitude is according to my current assessment of what it ought rationally to be. When I say who I believe the president is, my answer is beholden to the outward-directed reasons, even if I don't have to deliberate about them or "make up my mind" in order to answer.

11 Boyle (2009), 138 f.

that this claim about first-person awareness should be rejected. I pointed out that there is a wide range of cases in which a person speaks with first-person authority about her state of mind even though she cannot "avow" it in Moran's sense. Here is one kind of case: Max has a spider phobia. He is terrified of spiders—spiders in general, and the one on the pillow beside him in particular. Max doesn't take this fear to be rational; he knows that he suffers from a phobia. Still, when he says, "Get that thing away from me; I'm really afraid of it!", he is speaking (or shouting) about his own state of mind with first-person authority.

In that example, Max's fear is, by his own lights, irrational. Here is another kind of case, one in which the attitude at issue is not irrational: I sometimes look after my friend Adam's dog, Sadie, when Adam is out of town. I don't view it as rationally incumbent upon me to be fond of Sadie. I cannot answer the question of whether I *am* fond of Sadie by addressing a question about whether I ought rationally to be. Nonetheless, I'm fond of Sadie, and I have no trouble speaking with first-person authority about my fondness for her. In the postscript to my book, I argued that once it becomes apparent that one may speak with first-person authority regardless of whether one is talking about a belief that is dictated by the conclusion of theoretical reflection, an irrational fear, or an attitude that is neither dictated by reasons nor irrational, it is hard to go on being convinced by Moran's account of first-person authority.

3. Boyle allows that Moran can seem to have overstated his point:

> There seem plainly to be kinds of mental states of which our knowledge is both non-observational *and* non-deliberative: not just sensations but, for instance, appetites (i.e., brute, unreasoned desires for things of a certain kind) and what might be called "recalcitrant attitudes" (e.g., feelings of anger that I know to be unjustified but cannot overcome.) ... [I]f his [Moran's] account does not apply to such knowledge, it is not clear how he can justify his claim to have described the fundamental form of self-knowledge, the one that "makes the difference" between first-person awareness and the kind of awareness we might have of the mental states of another person.[11]

Boyle aims to show that, even so, there is an important sense in which Moran *has* explained the fundamental form of self-knowledge. The important point, according to Boyle, is not that Moran has managed to explain the only kind of first-person authority that there is. It is, rather, that he's accounted for the *fundamental* kind, where this is understood as the kind that the other kind(s) presuppose. According to Boyle, the real value of Moran's book lies not in extending the transparency point far beyond belief, or beyond belief at all. It lies in allowing us to see something essential to our agency—in helping us to understand a kind of authoritative self-awareness that is tied up with our nature as reflective agents.

Boyle accuses me (and others) of simply assuming that first-person authority is a unitary phenomenon—of presupposing that "a satisfactory account of our self-knowledge should be fundamentally uniform, explaining all cases of 'first-person authority' in the same basic way".[12] He asks why there shouldn't be a *kind* of first-person authority that is in place only when the subject's attitude is determined by her assessment of what the appropriate reasons require. Why, moreover, shouldn't we think that Moran has provided the right explanation of this kind of first-person authority? This, I think, is a good question—a reasonable response to the objection that I raised in the postscript to my book.[13] To address it, we need to ask whether Moran offers a satisfactory explanation of self-knowledge in the limited range of cases where a person's attitude *is* determined by her assessment of what the reasons require her to believe, desire, or fear.

According to Moran, when I come to the conclusion that, e.g., fear is rationally required of me, I am able to, as it were, read it off this conclusion that I am afraid. Putting his point this way suggests the following question: What entitles me to hold, on any particular occasion, that my actual attitude toward some object or proposition is the one that, according to me, is rationally required? After all (as we have seen), I don't *always* have the attitude that I think is rationally required of me. Imagine that I think through my current situation and conclude that I ought to be afraid. Am I afraid? If I simply avow the conclusion of my rational deliberation *as* my state of mind, is my avowal justified?

12 Boyle (2009), 141.

13 Boyle goes on to argue that first-person authority *cannot* be understood as a unitary phenomenon with a single explanation. I take him to be wrong about this, but I won't be directly engaging with his argument here.

Moran answers (what amounts to) this question in a paper published two years after his book. There he writes:

> [O]ne of the challenges to the Transparency claim can be put in the following way: What right have I to think that my reflection on the reasons in favor of P (which is one subject-matter) has anything to do with the question of what my actual belief about P is (which is quite another subject matter)? Without a reply to this challenge, I don't have any right to answer the question that asks what my belief is by reflection on the reasons in favor of an answer concerning the state of the weather. And then my thought at this point is: I would have a right to assume that my reflection on the reasons in favor of rain provided me with an answer to the question of what my belief about the rain is, if I could assume that what my belief here is was something determined by the conclusion of my reflection on those reasons. An assumption of this sort would provide the right sort of link between the two questions.[14]

Having come to the conclusion that my attitude ought rationally to be φ, what right do I have to claim that φ is, in fact, my attitude? Moran's answer is that I am entitled to *assume* that this is so; I'm entitled to assume that the attitude I in fact have is the one that, by my lights, the reasons call for me to have. In what follows, I shall refer to such an assumption as an RA (for "rationality assumption"). And in what remains of §3, I mean to question the way in which RAs figure in Moran's picture of self-awareness.

Before proceeding, I should perhaps say something about what I don't mean to be questioning in this area of Moran's thought. Here, I'd like you to consider the way in which the passage from Moran (2003) that I quoted in the preceding paragraph continues:

14 Moran (2003), 405. Boyle quotes this passage from Moran and glosses it as follows: "If a subject who possessed the concept of belief were entitled to assume that, in reaching the conclusion that *p* is true, he was coming to believe that *p*, then it seems that he could justifiably answer the question whether he believed that *p* by reflecting on ground for taking *p* to be true" (Boyle (2009), 137–8). In a footnote to this gloss, Boyle makes explicit that such assumptions won't *always* turn out to be true: "This is not to suggest that my merely taking myself to have a certain belief must make it so. As Moran emphasizes, his view does not demand that a subject be incorrigible about her own attitudes, or that her claims about her own attitudes have special authority no matter what their basis. What is important is that the question of what attitude I hold *can* often enough be settled by me on the basis of deliberation about whether *p*, and that—according to Moran—it is fundamental to the very possibility of thought about such attitudes that this should be so" (Boyle (2009), 137).

> And now let's ask, don't I make just this assumption, whenever I'm in the process thinking my way to a conclusion about some matter? I don't normally think that my assessment of the reasons in favor of P might have nothing to do with what my actual belief about P is, and it's hard to imagine what my thinking would be like if I did normally take this to be an open question. And if I did think that my actual belief about the rain might be left quite untouched by my reflections on the weather-related reasons, what do I imagine could possibly close this gap for me?[15]

I don't want to question the claim that, insofar as I take myself to be deliberating, I must assume—as I am "in the process of thinking my way to a conclusion"—that my assessments of the reasons for and against, e.g., believing that it's about to rain (or wanting to travel in China, or fearing that I'll lose my job) bear on what my attitude will turn out to be. (Perhaps it's a transcendental condition on my being a rational agent that I make this assumption.) But Moran is saying more than this. Again, he is claiming that I can and do self-ascribe attitudes by: (1) reasoning about what my attitude ought to be and (2) avowing the conclusion of this reasoning as what my attitude *is*. And as we've seen, a worry arises at this point: Sometimes my actual attitude doesn't jibe with the conclusion of my reasoning; what gives me the right to make any claim about what my attitude is, if I consider only what it ought to be? What entitles me to answer a question about my actual attitude by addressing a question about what my attitude should be? Moran invokes RAs at *this* point, to address *this* worry. When I say what my attitude is, I am assuming—and I'm entitled to assume—that this is *not* one of those cases in which there is, as it were, a gap between my assessment of what attitude is called for and my actual attitude.

Now, there are various concerns one might have about this position. My chief concern is this: It seems to me that insofar as someone needs to make this sort of assumption in order to self-ascribe an attitude, one is *alienated* from the attitude. This is connected to the fact that we don't often find ourselves saying such things as: "Given that my current desire is the one that I take to be most reasonable, I don't want to order any desert". The problem here is not merely that this would sound odd if I said it to a waitress; it's that such an utterance would mischaracterize the situation that I'm in (at least ordinarily) when I say what I want.

There *are* situations in which someone assumes he has an attitude that jibes with the reasons as they strike him. Consider the following:

15 Moran (2003), 405 f.

> Henry is a Navy SEAL, about to embark on a hazardous mission. His wife, who has gathered that he's being sent off to do *something* dangerous, asks him, "Are you afraid?". He answers, in a flat tone of voice, "No". She says: "Come on, Henry; you can tell me what you're feeling. We've talked about how dangerous your work is. Surely, you feel *some* fear". Henry thinks about what he's liable to be doing for the next few days and says, "Well, I see that there are lots of good reasons for me to be afraid, so I suppose that at some level, I must be".

Suppose that Henry is not lying to his wife in order to appear brave; he is trying, anyway, to speak honestly. Suppose, moreover, that Henry's self-ascription is correct; he *is* afraid at some level. Then, this is a case in which someone is able to correctly ascribe an attitude to himself only thanks to an RA. Insofar as this is Henry's situation, he is alienated from his own state of mind.[16] This is connected to the fact that when Henry finally ascribes fear to himself, he does not speak with (much[17]) first-person authority.

Contrast Henry's case with this one:

> It's 7 pm on a Monday. I'm at my office, working, when a friend phones to see if I want to go with her to a movie. I need only engage in a moment or two of practical reasoning—I think about the class that I have to teach in the morning, the letter of recommendation that's two days overdue, and the dissertation that I still haven't finished reading—whereupon I say to my friend, "I'm sorry; I'm sure I'd enjoy the movie, but I want to stay at the office this evening".

This is the kind of case that Moran's account should be best suited to explaining, one in which a subject rationally deliberates about what he wants to do, arrives at a single conclusion, and avows it. But now: at the point in the story when I tell my friend that I want to stay at the office, am I *assuming* that what I, in fact, want is in line with my assessment of what I ought to want? Surely not. At that moment, I *know* what I, in fact, want; my desire is (unlike Henry's fear in the last example) fully conscious, and I have no need of any such assumption. Indeed, it would be *more* plausible to claim that, in the case described, I know that my desire accords with my reasoning in part *because* I know that

16 Indeed, to the extent that this is his situation, the attitude that Henry self-ascribes is—to that extent—*unconscious*. (I argue that consciousness and first-person authority should be understood to come in degrees in Finkelstein (2003), §5.5.)

17 See the preceding note.

what I want is to stay at the office, than to claim, with Moran, that I can say what I want thanks to my assuming I have a desire that is dictated by the conclusion of my reasoning.

The same point obtains if we consider a case in which a belief, rather than a desire or fear, is self-ascribed:

> After painstakingly going over the evidence, Jill comes to the conclusion that her business partner, whom she took to be a friend, has been embezzling money from their shared company. She says, "I believe that Joel has been embezzling money from the company".

Unless we somehow flesh out the story in such a way as to make it plausible that Jill is talking about a belief of hers from which she's alienated, it's not going to make sense to claim that her self-ascription depends upon her assuming that she has an attitude that accords with the conclusion of her theoretical reasoning. If the belief that she's self-ascribing is an ordinary conscious one, then she has no need for such an assumption. (Here too, it would be more plausible to claim that she knows her attitude accords with her reasoning *because* she knows what she believes about Joel than it is to claim that she can say what she believes about Joel only because she is assuming that she has an attitude that jibes with her theoretical reasoning.) Whether we are talking about fears, desires, or beliefs, if a person needs an RA in order to avow her attitude, then she is alienated from it; she lacks "genuine first-person awareness" of it. And what Moran means to be offering is precisely an account of unalienated self-awareness.[18]

I'll sum up. Moran (2001) says that we can answer questions about our own attitudes by addressing outward-directed questions about what ought rationally to be believed, desired, or feared, given our situation. Moran (2003) adds that in so answering a question about his own attitude, a subject must assume that it jibes with his assessment of what his attitude ought rationally to be. I have been arguing that self-awareness

18 A similar objection might be raised in connection with detectivism. We do *sometimes* detect our own attitudes and emotions via self-observation (or inference or testimony). What the detectivist misses is that insofar as one must self-ascribe a mental state on this sort of basis, one is alienated from it. Detectivism goes wrong by representing the best cases of self-knowledge as if they were cases of alienated self-knowledge. I'm claiming that Moran winds up doing this as well.

does not ordinarily require such an assumption and that where it does, the subject is alienated from his own state of mind.[19]

4. The argument against Moran's position set out in the preceding section does not tell against Evans's original proposal for how we answer questions about our own beliefs. Moran is driven to represent avowals as depending on RAs because he shifts his starting point from (what I called in §2) TC to TC*. This allows him to extend an account of self-knowledge and first-person authority beyond belief—to desire, fear, intention, etc.—but it comes at the cost of claiming that, even in the best kind of case, a subject can only assume that the attitude he in fact has is one that accords with his assessment of what his attitude ought rationally to be. TC requires no such assumption; it makes no reference to reasons or rationality. According to Evans, I can answer a question about what I believe by addressing a corresponding question about what is *true*. I needn't address the question about what's true in a way that is, even by my own lights, rational. So I am able, e.g., to answer the question of whether I believe that God created the Earth by addressing the question "Did God create the Earth?"—and I can do this regardless of whether I answer the latter question on the basis of what I take to be good reasons.

Let's say that you are convinced by §3's objection to Moran's position. So you are not going to accept that (unalienated) self-knowledge and first-person authority depend upon RAs. Let's say, as well, that you are moved by Boyle's suggestion that we shouldn't just assume that

19 It might be possible to work out a variation on Moran's position that doesn't require the problematic appeal to RAs—by embracing a kind of disjunctivism about self-knowledge. In what we can think of as "the good sort of case", when deliberation results in the appropriate rational attitude, there is no gap between the outward-directed question "Ought I to be φ?" and the inward-directed "Am I φ?". Thus, in a good case, the subject is entitled—*without* an RA—to answer, e.g., the question, "Am I afraid?" by addressing the question "Ought I to be afraid?". Here, there is no need for a rationality assumption to bridge a gap between the two questions because *in a good case* there is no gap. If this sort of position could be made out, then it would be immune to the objection I've raised in this section. Although I won't work through the details, the right response to such a position, would, I believe, be a variation on what I'll say about Evans in §4. Evans's position is (already) like a disjunctivist version of Moran's in that there is no gap between the outward-directed question ("Is *p* true?") and the inward-directed question ("Do I believe that *p*?").

there is only one *kind* of first-person authority or (unalienated) self-knowledge. At this point, you might be tempted to draw a conclusion that could be put as follows: "Evans was right about the authority with which we speak about our own beliefs—but this authority turns out to be *sui generis*. We shouldn't try to explain our knowledge of our own desires, fears, wishes, etc. in the same way that we explain our knowledge of our own beliefs. We don't have the same *kind* of authority about our own desires, fears, and wishes as we do about our beliefs." In what remains of this essay, I'll argue that this is not the conclusion you ought to draw. I'll suggest that, rightly understood, TC leads—not to a picture of first-person authority as a divided phenomenon, but—to a unitary expressivist[20] account of the authority with which we speak about our own beliefs, desires, fears, moods, and sensations.

Just after the passage from *Varieties of Reference* that I quoted at the start of §2, Evans writes: "We can encapsulate this procedure for answering questions about what one believes in the following simple rule: whenever you are in a position to assert that *p*, you are *ipso facto* in a position to assert 'I believe that *p*'".[21] How is it that this "simple rule" is supposed to show that we needn't posit any sort of inward-directed observation or detection in order to make sense of belief self-ascriptions? An answer to this question could, I take it, be put as follows: We feel no pressure to posit anything like an inner sense in order to understand how it's possible for someone to address an outward-directed question, e.g., a question like "Is it about to rain?". Since one need do nothing *more* in order to answer a question like "Do you *believe* that it's about to rain?", we should feel no pressure to posit an inner sense in order to make sense of self-ascriptions of belief. Now, I'd

20 In earlier writings, I reserved the word "expressivist" for views according to which psychological self-ascriptions express a speaker's state of mind but are not themselves truth-apt. The position I argued for—according to which a speaker can express his state of mind by asserting truly that he is in it—did not, therefore, count as *expressivist*. But people were always inclined to describe me as a *kind* of expressivist, and since Bar-On (2004) distinguished between "simple expressivism" (according to which avowals are expressions that are not truth-apt) and "neo-expressivism" (according to which avowals are truth-apt expressions), I've given up resisting the word "expressivist" as a label for, among other things, the sort of position I've defended since Finkelstein (1994).

21 Evans (1982), 225.

like you to consider an objection to this line of thought. Imagine a detectivist who says the following:

> Suppose that someone asks you whether you believe that it's about to rain. You walk over to a nearby window, peer at the sky, and come to some conclusion; you form a judgment concerning whether or not it is about to rain. According to Evans, what is your "procedure" supposed to be at this point? (You haven't yet answered your interlocutor.) Presumably, the idea is this: If you have judged that it *is* about to rain (and so are "in a position to assert that *p*"), then you are to utter the sentence "I believe that it's about to rain" (or some suitable equivalent). If, on the other hand, you've judged that it's *not* about to rain, then you're to say instead, "I don't believe that it's about to rain". But now: how are you supposed to know the content of the judgment you've reached concerning whether it's about to rain? In order to follow Evans's "simple rule", you must have some way of finding out what judgment (if any) you have made about the outward-directed question. Thus, we still need to posit an inner sense—one that informs the subject of his own judgments—in order to understand how someone can say what he believes about, e.g., the weather. Really, all that Evans provides is a "procedure" by which a person can learn what she believes about some subject matter, *given* that she already (somehow) knows what she judges about it. And this is to offer precious little.[22]

I believe that we can make significant progress toward understanding first-person authority—not only about self-ascriptions of *belief*, but about a wide range of self-attributions—by thinking about how one might reply to this objection. A first pass at such a reply might go: What Evans means to be suggesting is *not* that you're entitled to utter a sentence of the form "I believe that *p*" just in case you have (somehow) learned a particular fact about yourself, viz., that you judge *p* to be true. It would be better to say that, according to Evans, we simply learn to use sentences of the form "*P*" and "I believe that *p*" *interchangeably*, at least to a certain extent. Thus, having just looked out your win-

22 A version of this objection could be raised against Moran's position. As we saw, according to Moran, I am able to say whether or not I believe that *p* (or fear *x*, or...) by coming to a deliberative conclusion about whether the reasons call for me to believe that *p* (or to fear *x*, or...). Our imagined detectivist would respond: "How, given this account of self-knowledge, are you supposed to know what deliberative conclusion (if any) you have reached on the question of whether the reasons call for you to believe that *p* (or to fear *x*)? Here too, we need to posit an inner sense—one that informs you of your own judgment about whether you ought rationally to believe that *p* (or to fear *x*). Without such a mechanism, you will be in no position to *say* how you have made up your mind—even if you have managed to make it up on the basis of the appropriate reasons."

dow at a fast-approaching bank of black clouds, you might say *either*, "It's about to rain", or, "I believe that it's about to rain". The latter assertion calls for no more inward observation than the former.

But suppose that our detectivist is unsatisfied with this reply to his objection. Imagine that he answers it as follows:

> You're failing to grasp the seriousness of my objection to Evans. You should think about a case in which a person is unaware of some judgment that he makes. Suppose that I unconsciously judge (perhaps on the basis of good, though painful, evidence) that my mother loves my brother more than she loves me. Thus I come to unconsciously believe that she loves him more. Now, it may be possible for me to gain knowledge of this belief of mine about my mother. But Evans doesn't suggest a way for me to do it. The interchangeability of "My mother loves my brother more" and "I believe that my mother loves my brother more" does me no good, for I am not in a position to say *either*. I can take advantage of Evans's "simple rule" only when my judgment concerning an outward-directed matter is *already conscious*—so only when my judgment that *p* is already something that I am (somehow) aware of. The postulation of an inner sense is needed in order to account for this sort of awareness. (So: my judgment that *p* is conscious just in case I learn about it via the inner sense.) The interchangeability of "*P*" and "I believe that *p*" simply does not account for the *difference* between: (1) a case in which someone can say what he believes, e.g., one in which you consciously believe that it's about to rain; and (2) a case in which someone is unable to say what he believes, e.g., one in which I unconsciously believe that my mother loves my brother more than she loves me.

Note that our imagined detectivist doesn't deny that "*P*" and "I believe that *p*" are, as it were, interchangeable. But he cannot see how this fact obviates the need for inner sense. (Indeed, what our detectivist is liable to take away from Evans is that inner sense is needed merely to look out a window and say, "It's raining".)

Now, while there *is* something to our detectivist's objection, he is nonetheless missing a point of crucial significance. As a step toward bringing this point into view, we might wonder about the interchangeability of "*P*" and "I believe that *p*". Why, in spite of their having different truth conditions, am I able to use (e.g.) "It's about to rain" and "I believe it's about to rain" so interchangeably? The answer could be put this way: It's because (in most circumstances) I'll express the same attitude—the same belief—regardless of which of these sentences I utter. If someone wants to know whether I believe that it's about to rain, I can give him what he wants by uttering *either* sentence. Notice that one upshot of this is that "I believe it's about to rain" may be used to express the very state of mind that it self-ascribes.

For all that Evans says, our imagined detectivist cannot see how self-ascriptions of belief are supposed to be possible unless the self-ascriber inwardly detects *either* her own beliefs *or* the judgments that correspond to them. I want to suggest that the crucial point missed by our detectivist just is that a statement like "I believe it's about to rain" may be understood as an expression of the speaker's state of mind—an expression of the speaker's belief that (and also her judgment that)[23] it is about to rain. Suppose that, upon looking out a window and concluding that it's about to rain, you feel disappointed and sad. (Perhaps it's your wedding day, and you've arranged for an outdoor reception.) You utter the sentence "I believe that it's about to rain", whereupon you collapse onto the floor and begin to cry. We should understand the relation between your belief and your self-ascription of belief as akin to that between your sadness and your crying—so: not as depending on inward detection, but as an expression, a manifestation, of your psychological condition.[24]

At the start of the present essay, I asked how it might be possible to reject detectivism without making it appear miraculous that we are able to easily and accurately self-ascribe beliefs, desires, fears, hopes, and the like. We have arrived at what I take to be a good answer (or, anyway,

---

23 I'm not suggesting anything controversial here about how to understand the relation between a belief that *p* and a judgment that *p*. The point is, rather, that an utterance like "I believe it's about to rain" may express *various* psychological states and goings-on: not only a belief about the weather and its corresponding judgment—however precisely we understand the relation between beliefs and judgments—but (e.g.) disappointment, resignation, or anger. Just as it's a mistake to hold that if something is an expression, it can't also be a truth-evaluable assertion, it's a mistake to hold that if something expresses a belief, it can't express anything else.

24 Because he doesn't see this point, our imagined detectivist (also) misunderstands the distinction between conscious judgments/beliefs and unconscious ones. What it means for a belief or a judgment to be conscious, rather than unconscious, is not that the subject is aware of it—or that she is aware of it via inner sense. Rather, a conscious belief or judgment (or desire or fear) is one that the subject is able to express in a particular way—by self-ascribing it. (I defend this claim in Finkelstein (1999) and Finkelstein (2003), §§5.4–5.5.) We can, after all, account for the difference between the two cases that our imagined detectivist adduces in his objection to Evans (the case in which someone consciously believes that it's raining and the case in which someone unconsciously believes that his mother loves his brother more than she loves him) without appealing to an inner sense.

the start of a good answer) to this question. We can reject detectivism—at least about belief self-ascriptions—by recognizing that saying, e.g., "I believe it's about to rain" is a way in which I can express (rather than report an observation of) my attitude about the weather.[25] Evans brings us close to this point by, in effect, claiming (correctly) that I no more need to look inward in order to say, "I believe that *p*", than I need to look inward in order to say, "*P*". But the significance of this claim comes into focus only when we attend to what explains it—viz., the fact that saying, "*P*", and saying, "I believe that *p*", are two (amongst various) ways in which I might *express* my belief that *p*.

Until we get a firm grip on the point that (as I put it a moment ago) Evans only brings us close to, we are likely to react to what he says about belief self-ascriptions in one or the other of two ways. On the one hand, we might think, as our imagined detectivist does, that Evans offers "precious little"—showing merely that once I have found out whether I *judge* that *p* is true, I can say whether I *believe* that *p* is true without any *further* inward observation. On the other hand, we might have a sense that Evans is onto something important, but—failing to get the point about expression into focus—fixate on the fact that "*P*" and "I believe that *p*" may be used in place of each other across a wide range of contexts. Then we're liable to think that in order to account for the authority with which we speak about attitudes other than belief, we should try to find outward-directed questions that are, for the subject, transparent to such questions as "Do I want to go to a movie this evening?". Once we've started down this road, we are bound to conclude that first-person authority has no single explanation—if only because we are not going to find outward-directed questions to which questions like "Am I feeling pain?" or "Am I in a good mood?" are transparent.

We can, after all, extend a non-detectivist understanding of belief self-ascriptions not only to other attitudes (regardless of whether or not they jibe with our assessments of what's rationally called for), but to moods and sensations as well. Just as I can express my belief by saying, "I believe…"; I can express my desire by saying, "I want…"; my hope

25 To express one's state of mind is not to report an observation of it; I take this to be a point of significance for the later Wittgenstein. When a child comes into language and learns, among other things, to express his states of mind by self-ascribing them, a great deal changes, but expression doesn't suddenly call for an epistemic basis. For more on Wittgenstein's treatment of this point, see Finkelstein (2001) and Finkelstein (2010).

by saying, "I hope…"; my elation by saying, "I'm elated"; and the pain I feel in my shoulder by saying, "I have a pain in my shoulder". Just as I needn't inwardly detect my belief about the weather in order to express it by saying, "I believe that it's about to rain", I needn't inwardly detect my irrational fear of spiders in order to express it by saying, "I'm afraid of spiders".

In §1, I characterized the "transparency approach"—according to which we are able to speak about our own states of mind by answering questions about the world outside us—as an approach *toward* understanding self-knowledge and first-person authority. I have, in effect, argued that this approach gets us only partway to its goal. I've also sketched what I take to be a more satisfactory way of addressing the question from which I began, the question of how we can reject detectivism without making our capacity to self-ascribe psychological states and events come to seem miraculous. Filling in this sketch would entail saying more than I have here about, among other things, how to understand the notion of expression. But, if you have followed me this far, then you'll understand why I think the best response to Evans (and Shoemaker and Moran) is to see what's right in the transparency approach as pointing in the direction of an expressivist one.

References

Bar-On (2004): Dorit Bar-On, *Speaking My Mind: Expression and Self-Knowledge*, Oxford.

Boyle (2009): Matthew Boyle, "Two Kinds of Self-Knowledge", in: *Philosophy and Phenomenological Research* 78 (1), 133–164.

Evans (1982): Gareth Evans, *The Varieties of Reference*, J. McDowell (ed.), Oxford.

Finkelstein (1994): David H. Finkelstein, *Speaking My Mind: First-Person Authority and Conscious Mentality*, Ph.D. Dissertation. University of Pittsburgh, Ann Arbor: ProQuest/UMI. (Publication No. 9521463.

Finkelstein (1999): David H. Finkelstein, "On the Distinction between Conscious and Unconscious States of Mind" in: *American Philosophical Quarterly* 36 (2), 79–100.

Finkelstein (2001): David H. Finkelstein, "Wittgenstein's 'Plan for the Treatment of Psychological Concepts'", in: T. McCarthy/S. Stidd (eds.), *Wittgenstein in America*, Oxford, 215–236.

Finkelstein (2003): David H. Finkelstein, *Expression and the Inner*, Cambridge/MA.

Finkelstein (2010): David H. Finkelstein, "Expression and Avowal", in: K. Jolley (ed.), *Wittgenstein: Key Concepts*, Durham/UK, 185–198.

Locke (1689/1975): John Locke, *An Essay Concerning Human Understanding*, P.H. Nidditch (ed.), Oxford.
Moran (2001): Richard Moran, *Authority and Estrangement: An Essay on Self-Knowledge*, Princeton.
Moran (2003): Richard Moran, "Responses to O'Brien and Shoemaker", in: *European Journal of Philosophy* 11 (3), 402–419.
Shoemaker (1994): Sydney Shoemaker, "Self-Knowledge and 'Inner Sense', Lecture II: The Broad Perceptual Model", in: *Philosophy and Phenomenological Research* 54 (2), 271–290.

On Finkelstein's Account of the Distinction between Conscious and Unconscious States of Mind[1]

James Doyle

1. Introduction

Some of my mental states are conscious – all my sensations, presumably, and many beliefs, hopes, fears, emotions, and so on. Others of my mental states, it seems, and I will assume, are unconscious. I may be consciously angry with my father, but I might be unconsciously so instead. An account of what it is for a mental state of mine to be conscious should make sense of certain marked dissimilarities between my relation to my own conscious states and my relation to other people's. For whether or not I can be mistaken about my own conscious states, I do not *seem* to know about them, when I do, on the basis of anything like evidence, and there *seems* to be something strained and implausible in the idea that I know about them from anything like observation. Whereas evidence and observation are of course very much to the point in my ascriptions of conscious states to other people, and theirs to me. This 'first-/third-person asymmetry' has led some philosophers to try to think of (at least some) first-person avowals of conscious mental states as first and foremost *expressions* of them, so that "I have a headache" expresses my pain in something like the way a groan would. This certainly gives us asymmetry: I can hardly express *someone else*'s pain in that sort of way. But this sort of account seems to miss out something else important: what the first-person avowal and third-person ascription have in common. For they *both* seem to be ascriptions: when you tell someone else that I have a headache, you seem to be ascribing the same state to me as I ascribe to myself when I say I have a headache; and both ascriptions seem to be made true by the same state of affairs,

1 I would like to thank Jim Conant, David Finkelstein, James Ladyman and participants in the University of Chicago Wittgenstein Workshop for helpful comments on earlier versions of this paper.