Transparent Minds: A Study of Self-Knowledge, by Jordi Fernandez. Oxford: OUP. 2012.

Reviewed by Matthew Parrott (forthcoming in European Journal of Philosophy)

(PENULTIMATE DRAFT)

Although the question of how we know our own minds has had a prominent place throughout the history of philosophy, a recent surge of interest in it has been sparked by an observation made by Gareth Evans, and heavily emphasized by Richard Moran. The thought is that, rather than introspecting or turning our attention toward our minds, we are typically able to answer questions about our own psychological attitudes simply by considering facts in the external world. For example, the standard way in which someone answers the question of whether or not they believe that the tulips in the garden are red is not by focusing on some set of mental states but by attending to the tulips. By doing this, it looks like one can come to know about one's *beliefs* simply by considering evidence or reasons that have nothing to do with one's psychology. This is exceptionally puzzling. For better or worse, this phenomenon has come to be called 'transparency' and it is at the center of Jordi Fernandez's *Transparent Minds.*

Several philosophers have written on transparency but it would be difficult to imagine someone doing a better job than Fernandez. *Transparent Minds* is exceptionally well written. Fernandez's prose is refreshingly straightforward and he articulates the central arguments for his views with marvelous clarity and precision. For these reasons alone, the book is a rare achievement and well worth reading.

The goal of the book is to explain two powerful intuitions many philosophers have about self-knowledge. First, it seems that the way we ordinarily know our own minds is special; it is fundamentally different from the way we know the minds of other people. As

Fernandez describes it, 'in normal circumstances, we do not acquire knowledge of our mental states by looking at ourselves from outside, as it were, and trying to make sense of our own behavior.' (pg. 6) In addition to being peculiar, however, this first-personal way of knowing is thought to be uniquely secure from an epistemic perspective. Thus, Fernandez says we tend to think that 'a subject's justification for her beliefs about her own mental states is stronger than anybody else's.' (pg. 7)

Because he wants to respect these intuitions, Fernandez's book focuses on epistemic properties of self-knowledge. Not every theory of self-knowledge takes this approach and even those that do differ in significant ways. In chapter 1, Fernandez helpfully distinguishes his account from several other leading views. He uses the chapter primarily to draw out six desiderata that he believes an adequate theory of self-knowledge should meet, in addition to explaining the previous two intuitions (pg. 38). The problem with the other accounts he considers is that each one fails in some way to meet his desiderata. For example, functionalist theories and so-called 'no-reasons' accounts are diagnosed as having 'trouble accommodating the intuition that our higher-order beliefs constitute a cognitive achievement' (pg. 28) and deliberative accounts, like Moran's, are said to imply that our way of acquiring self-knowledge 'should be infallible, which it is not.' (pg. 21)

There is room to question Fernandez's objections to these other views. For instance, it is not clear to me why the deliberative view entails infallibility. Surely the view could allow for a subject to be mistaken about her attitudes when they are not formed via reflective deliberation and this seems sufficient to meet the desideratum of allowing 'for the possibility that self-attributions of mental states are wrong.' (pg. 38) Perhaps one could not be wrong when it comes to attitudes that are actually formed through deliberation, perhaps those are self-intimating, but that is not the same as infallibility. One could still be liable to error in

cases where one fails to deliberatively form an attitude and, contrary to what Fernandez claims, one can deliberate on a question like whether or not the tulips are red without thereby coming to any settled belief on the matter. One might fail to accomplish the goal of one's deliberation without thereby failing to deliberate, just like one might fail to run a mile without failing to run. If the deliberative view does not entail infallibility, however, it may turn out to be a viable contender to Fernandez's own account. Indeed, Fernandez's theory is notably similar to Moran's in that they are both motivated primarily by the phenomenon of transparency. Fernandez names his theory the 'bypass model' and he often emphasizes that it will 'naturally rest on Evans's observation' (pg. 50).

The core of Fernandez's bypass model is developed over the course of chapters 2 and 3 and it is worth noting that Fernandez deliberately restricts the account to self-knowledge of beliefs and desires. Those who insist on having a general theory of self-knowledge may be uncomfortable with this strategy. But since there are reasons to think no such theory will be forthcoming, and since the bypass view offers a credible explanation of the way in which we know our judgment-sensitive attitudes, it deserves serious consideration.

The central claim of the bypass view is that we form beliefs about our own beliefs and desires on the basis of our grounds for those beliefs and desires. For example, when I believe that I believe that the tulips are red, this is normally based on the same grounds I have for believing that the tulips are red, which, according to Fernandez, is my perceptual experience of the tulips. In order to base my higher-order belief on my experience, two things must be true. First, the experience of red tulips must actually cause the higher-order belief that I believe that the tulips are red and, second, I must be disposed to believe that I am having the experience. Thus, on Fernandez's picture, basing one's higher-order beliefs on

an experience is primarily a causal notion. One does not need to think one's higher-order beliefs are justified by experiences nor indeed have any thoughts about the epistemic role of one's experiences. As Fernandez claims, 'forming a belief on the basis of some states does not require believing that one occupies that state, and it does not require believing that, if one is in that state, then the content of the belief being formed is likely to be the case.' (pg. 56) Supposing this is right, how does my experience of red tulips justify my belief about what I believe?

Fernandez's answer relies heavily on an externalist conception of epistemic justification, more specifically on the reliability of certain causal relations. Put simply, he believes that an experience constitutes sufficient justification for a belief just in case an experience of that kind 'tends to correlate with the type of state of affairs that makes the belief true.' (pg. 43) This is not so implausible when it comes to one's perceptual belief that the tulips are red. In this case, the thought would be that my experience justifies my belief that the tulips are red because it is a type of experience that is reliably correlated with what makes that belief true, namely the tulips being red.

But, we might think the same picture is much less plausible when it comes to my beliefs about what I believe. Indeed, part of the reason the phenomenon of transparency seems so puzzling is that it is very obscure how an experience *of tulips* could possibly provide someone with an epistemic reason for a belief about her beliefs. Fernandez hopes to dissolve this sense of puzzlement by emphasizing that the source of justification, just as in the perceptual case, is the reliability of the causal connection between my experience of red tulips and what makes my higher-order belief true, namely my believing that the tulips are red. Since my experience of red tulips does tend to cause me to believe that the tulips are red, it is able to justify the higher-order belief that I believe that the tulips are red. Because of

this correlation between my perceptual experience and what makes my higher-order belief true, my experience of red tulips is supposed to be 'the kind of state that could provide a rational basis for the self-attribution of a mental state.' (pg. 52)

However, from a subject's point of view, it still seems confusing how the experience of red tulips could *rationalize* my belief that I believe the tulips are red. It might do so if I knew that experiences of that kind reliably caused beliefs about tulips. But if I don't know about the reliability of that correlation, how could it rationalize the sort of behavior Evans comments on? How could it make sense for me to *choose* to answer your question about whether or not I believe the tulips are red by focusing on the tulips? Reliability looks like it has a chance at justifying a subject's perceptual beliefs because all the subject has to do is something she cannot really help doing; she need only take the content of her experiences at face value. But it looks like something more must be done to acquire self-knowledge, especially if one is answering an explicit question about one's beliefs, which, Fernandez claims 'requires effort, since it requires that we direct our attention at the world' (pg. 62). The question for the bypass model is what reason someone could have for intentionally making such an effort.

This difficulty arises in part because, as becomes especially clear in chapter 4, the bypass view allows any mental state to count as adequate justification for a higher-order belief as long as it stands in the right causal relation to that belief's truth conditions. So if my perceptual experience of the red tulips regularly causes a feeling of tranquility, the view implies that I would be justified in believing that I feel tranquil on the basis of seeing red tulips. Although I may not actually be disposed to form beliefs in these ways, it seems that I could. And Fernandez seems committed to the conclusion that were I to do this, my beliefs would be adequately justified. This would result, I think, in extremely bizarre forms of

transparency. Do we really want to say that I am justified in deciding to answer a question about whether or not I feel tranquil by attending to nothing but the color of tulips? Could any causal relation make this sort of behavior intelligible, much less rational?

Fernandez's reliance on causal relations also raises questions about his proposal for explaining how self-knowledge enjoys strong epistemic justification. It seems to me that, like many philosophers writing on self-knowledge, Fernandez's intuition about the comparative epistemic strength of self-knowledge is largely due to his taking for granted a particular conception of the way we know others' minds. He writes, 'in order for you to be justified in believing that I have a belief, you typically need to observe my behavior (including my verbal behavior) and infer from it that I have the belief in question as the best explanation of your observations. There are some aspects of this procedure that make you liable to error in ways in which I am not.' (pp. 57-58) In fact, there are two. First, there is the possibility of a perceptual error and, second, there is the possibility of making an inferential mistake. According to the picture Fernandez adopts, your way of knowing about my attitudes is reliable only because these two errors do not typically occur. Nevertheless, because it is susceptible to these errors, your way of knowing about my mind is supposed to have a somewhat weaker epistemic status than my self-knowledge.

This characterization of the way we know others' minds is not uncontroversial. Several philosophers resist the notion we acquire knowledge of others' minds through inference. Rather, an alternative proposal is that we know the mental states of others on the basis of perception. On a perceptual view, however, it is questionable whether the bypass model really secures stronger epistemic justification for self-knowledge. Suppose you know my mental states on the basis of perception and I know them, as Fernandez claims, on the basis of their grounds. In one sense, you would be liable to a type of error that I am not, a

perceptual error. However, this is not fundamentally different from the kind of error that I am vulnerable to. Just as perceptually based knowledge rests on a causal relation, my self-knowledge acquired through bypass rests on a causal relation. It *might* be true that causal breakdowns are more common in perception but it might not be. If we compare two possible worlds, one where perceptual systems suffer causal failures but the mechanisms underlying self-knowledge function fine and the other where the reverse is true, it is not obvious to me which world is closer.

Fernandez seems to assume that perceptually based beliefs are epistemically less safe than those not based on perception. But we need not accept this, especially not if self-knowledge also rests on a causal mechanism. Of course, since it does not require any sense organs, there remains a sense in which my beliefs about my mental states are not vulnerable to certain kinds of errors, but this is just another way of saying that my way of knowing is distinctive. Much more would need to be said before concluding that it enjoys stronger epistemic support.

In the final three chapters, Fernandez applies the bypass model to explain Moore's paradox, the delusion of thought insertion, and self-deception. Although space prohibits commenting on each of these topics, I was slightly bewildered by one aspect of his discussion of thought insertion in chapter 5. Fernandez initially describes thought insertion as a 'psychological disorder wherein the subject is under the impression that certain thoughts that she has are not her own thoughts' and presents seven actual cases that nicely illustrate this delusion. (pg. 141) However, he then explicitly assumes that, 'the mental states disowned by patients who suffer from thought insertion...are beliefs.' (pg. 146) This makes it easy to explain thought insertion in terms of the bypass model but it seems to mischaracterize the phenomenon. Many reports from individuals who have experienced thought insertion do

not appear to cite beliefs. Indeed, in two of Fernandez's cases, subjects seem to report commands being inserted into their mind and in four of the remaining five there is no clear indication as to what type of mental state is involved. Fernandez's stated motivation for making the assumption that a schizophrenic's experience involves inserted beliefs is that otherwise 'it is hard to see why that experience would seem odd to her.' (pg. 146) Indeed. But that is precisely why the delusion is so perplexing.

Nevertheless, Fernandez deserves considerable praise for recognizing that an adequate theory of self-knowledge should shed light on anomalous ways in which subjects relate to their mental states. Fernandez never loses sight of how self-knowledge can be fragile and how, at times, it can be extraordinarily difficult to hold on to. Because of this, *Transparent Minds* is always thoughtfully attentive to our actual knowledge of our own minds.