HOW WE KNOW WHAT WE THINK

Quassim Cassam

University of Warwick

Abstract

Assuming that knowledge of our own beliefs is usually epistemically and psychologically immediate a natural question is: how is such immediate self-knowledge possible? I examine and criticize Richard Moran's response to this question and develop a different account. My alternative draws on the idea that immediate self-knowledge results from the operation of a sub-personal monitoring mechanism. I express doubts about the extent to which knowledge of our own beliefs is immediate, and suggest that some versions of the immediacy intuition rest on a confusion between belief and judgement.

1. Immediate Self-Knowledge

The immediacy intuition about self-knowledge is that knowledge of our own beliefs and other propositional attitudes is usually immediate.[1] If immediate knowledge is defined as knowledge that is not based on observation, evidence or inference then the immediacy of self-knowledge can seem puzzling.[2] For, on the one hand, knowledge of our own beliefs is knowledge of contingent matters of fact. For example, I believe that Quine was born in Akron and I know that this is what believe. However, it is a contingent fact about me that I have this belief, and it is easy to conceive of my not having acquired it. On the other hand, the usual presumption is that knowledge of contingent matters of fact must be based on observation, inference or evidence. How, then, is it possible for me to know that I believe that Quine was born in Akron other than on the basis of observation, inference or

evidence?[3] There are examples of immediate knowledge of contingents facts but self-knowledge is not relevantly similar to such examples.[4]

In this paper I discuss one prominent attempt to explain the immediacy of self-knowledge in terms of the related notions of transparency and avowal. This is the explanation proposed by Richard Moran in his book Authority and Estrangement. I will draw attention to some difficulties with Moran's account and then go on to suggest another way of explaining how knowledge of our own beliefs can be immediate. However, I also contend that there is a question about the respectability of the immediacy intuition, and that at least some alleged examples of immediate knowledge of our own beliefs are in fact examples of immediate knowledge of our own judgements.

According to Moran, when the question arises what my belief about something is, avowal is a way of answering the question and hence a way of coming to know it. This knowledge is not based on inference, evidence or observation so the idea that we can know our thoughts or beliefs by avowing them accounts for the immediacy of much of our self-knowledge. Avowal is a way of answering a question about one's beliefs that obeys the 'Transparency Condition'. This condition states that:

> Ordinarily, if a person asks himself the question "Do I believe that P?", he will treat this much as he would a corresponding question that does not refer to him at all, namely, the question "Is P true?"' (2001: 60).

Moran's way of putting this is to say that the question "Do I believe that P?" is transparent to the question "Is P true?". Specifically, 'a first-person present-tense question about one's belief is answered by reference to (or consideration of) the same reasons that would justify an answer to the corresponding question about the world' (2001: 62).

I argue in part 2 that this approach fails to account for the immediacy of self-knowledge, whether immediacy is understood as a psychological or as an epistemic notion. In brief, the point is that while consideration of the reasons in favour P might lead one to judge that P, judging that P is not the same as believing that P and does not ensure that one believes that P. This need not prevent one from knowing that one believes that P on the basis of one's knowledge or awareness that one judges that P, but the resulting knowledge of one's belief is not immediate.

This leaves us with two options: (a) question the immediacy intuition or (b) endorse the intuition while looking for a different explanation of the possibility of immediate self-knowledge. These options are explored in part 3. It seems, on the one hand, that the extent to which knowledge of our own beliefs is immediate is often exaggerated. One reason is that writers on this topic often fail to distinguish clearly between the idea that knowledge of our own mental acts of judging or thinking is immediate and the far more contentious idea that knowledge of our standing mental states like belief is immediate. On the other hand, it is difficult to accept that there is nothing to the immediacy intuition. I believe my name is Quassim Cassam and I know that this is what I believe. My knowledge in this case appears to be both psychologically and epistemically immediate, and this needs to be accounted for. If Moran fails to account for it, then we need a better account. I will conclude, in part 4, by briefly indicating what such an account might look like.

2. The Limits of Avowal

Imagine this: I am trying to persuade the administration of my university to make extra funds available to my department. I have been promised the support of my colleague Dr. Nogood but he lets me down. I am disappointed but not surprised. Asked why not I say

that I have always believed that Nogood would let me down in a situation like this. Suppose that the question arises whether this is really something that I have always believed, that is, believed since I first met Dr. Nogood. What is the role of the Transparency Condition in relation to this question? Let P be the proposition that Nogood would let me down in a situation such as the one in which I now find myself. The issue is whether I can answer the question "Have I always believed that P?" by considering the reasons in favour of P itself. A strong consideration in favour of P is that Nogood has just let me down but it is hard to see how this puts me in a position to know that I have always believed that P. If I have always believed that P then it must be true that I did believe that P at some point in the past but the newly acquired reason in favour of P – the fact that Nogood has just let me down - has no bearing on whether I believed in the past that he would let me down. I might have believed that P in the past even though the reasons in favour of P now strike me as weak, and I might have failed to believe that P in the past even though the reasons in favour of P now strike me as strong. In the past, these same reasons might not have been available to me or my view of them might have been different.

It might be objected that the Transparency Condition is not designed to account for our knowledge of our past beliefs. The proposal is that a first-person present-tense question about one's belief is to be answered by reference to the same reasons that would justify an answer to the corresponding question about the world. Since the question "Have I always believed that P?" is not one of the requisite form one should not be surprised that it is not covered by Moran's account. It is worth noticing, however, that the question "Do I believe that P?" can itself be read in a way that is problematic for Moran's purposes. For, as Shah and Velleman observe, this question 'can mean either "Do I already believe that P (that is,

antecedently to considering this question?" or "Do I now believe that P? (that is, now that I am considering the question?"' (2005: 506). While the latter question can be answered by considering the reasons in favour of P and forming a belief with respect to this proposition 'one cannot answer the question whether I already believe that P in a way that begins with forming the belief' (ibid.). This is the problem of antecedent belief. In essence, the problem is that it does not appear to be possible to come to know one's antecedent beliefs by now avowing them.[5]

Next, consider a case in which the question "Do I believe that P?" is read as asking not whether I already believe that P but whether I believe that P now that I am considering the question. Isn't plausible, at least in this case, that I can answer by consideration of the reasons in favour of P itself'? One problem with this is that I might be convinced that P and be unable to shake off this conviction even though I recognize that the reasons in favour of P are weak. In this case the belief that P perseveres or sticks despite the acknowledged weakness of the evidence in favour of it.[6] The reverse of this is also possible: the reasons in favour of P might be strong enough to get one to judge that P but one still fails to believe that P. Peacocke gives a nice example of this:

> Someone may judge that undergraduate degrees from countries other than their own are of an equal standard to her own, and excellent reasons may be operative in her assertions to that effect. All the same, it may be quite clear, in decisions she makes on hiring, or in making recommendations, that she does not really have this belief at all (1998: 90).

So we now have two versions of what might be called the sticking problem for Moran's account. In the first version the belief that P sticks despite the acknowledged absence of

good reasons in favour of P. In the second version, the belief that P fails to stick despite the acknowledged presence of good reasons in favour of P. Either way, consideration of the reasons in favour of P itself fails to settle the question whether one believes that P.

What is the precise significance of the sticking problem for Moran's account of self-knowledge? Here are two features of sticking scenarios that are worth noting:

(1) They show that what one judges to be the case and what one believes to be the case might fail to coincide. One can judge that P and still fail to believe that P, and one can believe that P even though one judges that not-P.

(2) They show that when the question arises whether one believes that P avowal is not the only way of answering the question. I can have good evidence that I believe that P, and come to know that I believe that P on the basis of this evidence, even if I am unwilling to judge that P.

The second of these points suggests that self-knowledge without transparency is possible but this is not something that Moran denies. He regards it as 'undeniable' that 'a person can find herself in a situation where the evidence in favour of attributing a belief to herself is stronger than the evidence she has for the truth of the belief itself' (2003: 407). In such cases, the question "What do I believe?" is answered in a 'theoretical' rather than a 'deliberative' spirit. One might still end up knowing that one believes that P but this will be knowledge 'by attribution' (2003: 410) rather than by avowal. Moran's point here is that attributional self-knowledge cannot be the most basic form of self-knowledge for a rational agent. For part of what it is to be such an agent 'is to be able to subject one's attitudes to review in a way that makes a difference to what one's attitude is' (2001: 64).

With regard to (1), let us return to the idea that I can answer the question "Do I believe that P?" by consideration of the reasons in favour of P. Suppose that consideration of the reasons in favour of P leads me to conclude that P. The first thing to notice is that concluding that P is not the same thing as believing that P. To conclude that P is to judge that P, and judgement is a mental act. The act of concluding or judging that P normally leads to the formation of the belief that P but, as Peacocke's example illustrates, is not guaranteed to do so. Even when judging P does lead to the formation of the belief that P the belief is formed via the judgement that P.[7] This is consistent with the idea that 'the goal of deliberation, whether theoretical or practical, is conviction' (Moran 2001: 131). The point is rather that, when all goes well, theoretical deliberation results in the conviction that P via the mental act of affirming that P in response to the reasons in favour of P.

What is the epistemological significance of this account of the relationship between judging and believing? Suppose that my recognition of the reasons in favour of P leads me to judge that P. Suppose, also, that it is taken for granted that I know that I judge that P.[8] What is the relationship between knowing that I judge that P and knowing that I believe that P? If judging that P were equivalent to believing that P then knowing that I judge that P would amount to knowing that I believe that P. However, it is false that judging that P is believing that P since, as we have seen, it is possible to judge that P without believing that P. Nevertheless, one might still think that knowing that I judge that P amounts to knowing that I believe that P the following sense: if I know that I judge that P, and I am entitled to assume that my judgements normally determine my beliefs, then I can conclude that I believe that P.

Improbable as this account of self-knowledge might sound, it is what is suggested by Moran's response to what he regards as a major challenge facing his theory. The challenge is to explain what right I have to think that 'my reflection on the reasons in favour of P (which is one subject-matter) has anything to do with the question of what my actual belief about P is (which is a quite different subject matter)' (2003: 405). He responds that I would have a right to assume this 'if I could assume that what my belief here is was something determined by the conclusion of my reflection on those reasons' (ibid.). He adds:

> And now, let's ask, don't I make just this assumption, whenever I'm in the process of thinking my way to a conclusion about some matter? I don't normally think that my assessment of the reasons in favour of P might have nothing to do with what my actual belief is, and it's hard to imagine what my thinking would be like if I did normally take this to be an open question (2003: 405-6).

The conclusion of my reflection on the reasons in favour of P is a judgement so assuming that my belief concerning P is determined by the conclusion of my reflection on the reasons in favour of P is equivalent to assuming that my belief concerning P is determined by whether I judge that P. Call this the linking assumption, since it concerns the link between what one judges and what one believes. It appears that the role of the linking assumption in Moran's account is to facilitate the transition from knowledge of what I judge to knowledge of what I believe. If I know that I judge that P then, given the linking assumption, I am in a position to know that I believe that P.

How does this bear on the supposed immediacy of the self-knowledge? In order to answer this question, we need to be clearer about the notion of immediate knowledge. On

one view, immediate knowledge is non-inferential knowledge.[9] At any rate, it is plausible that immediate knowledge must at least be non-inferential even if it is not just equivalent to non-inferential knowledge. Suppose, also, that there is a justification condition on knowing, so that I know that P only if I am epistemically justified in believing that P. Then my knowledge that P is non-inferential if and only if my justification for believing that P is non-inferential. Finally, for my justification for believing that P to be non-inferential it must not come, even in part, from my having justification to believe other, supporting propositions.[10]

Given that immediate knowledge must be non-inferential, one way of showing that my knowledge that I believe that P is not immediate is to show that it is inferential. It is worth noting, however, that the notion of immediacy can either be understood epistemically or psychologically. Being non-inferentially justified in believing that P is a necesssary and arguably sufficient condition for one's knowledge that P to be epistemically immediate. In contrast, one's knowledge that P is psychologically immediate just if one did not arrive at the belief that P by reasoning or by inferring P from other propositions which one believes.[11] In principle, a person's knowledge that P could be psychologically immediate without also being epistemically immediate: just because I did not come to know that P by reasoning or inference it does not follow that my justification for believing that P is non-inferential.

The distinction between epistemic and psychological immediacy leaves us with two questions:

(i) When I come to know that I believe that P by following the transparency procedure is my knowledge epistemically immediate?

(ii) When I come to know that I believe that P by following the transparency procedure is my knowledge psychologically immediate?

Given the role of the linking assumption in Moran's account it appears that the answer to both questions is 'no'. For when I come to know that I believe that P by following the transparency procedure, my justification for believing that I believe that P comes, at least in part, from my having justification for believing the linking assumption.[12] This makes my knowledge that I believe that P inferential, and inferential knowledge is not epistemically immediate. It is also not psychologically immediate since coming to know that I believe that P on the basis of the transparency procedure looks like a clear-cut case of coming to know what I believe by reasoning.[13]

One response to this argument might be something along the following lines: just because I have to make the linking assumption whenever I am in the process of thinking my way to the conclusion that P, it does not follow that my justification for believing that I believe that P comes from my having justification to believe the linking assumption. As long as this assumption is not the source of my justification for believing that I believe that P, there is no reason to suspect that my justification is inferential or that my knowledge I believe that P is anything other than epistemically immediate. The problem with this, however, is that it is not clear what work the linking assumption is supposed to be doing if it is not seen as contributing in any way to my being justified in believing that I believe that P. It is presumably not enough for Moran's purposes that the linking assumption is correct. It is an assumption I must actually make for me to have the right to think that my reflection on the reasons in favour of P has something to do with the question of what my actual belief about P is. But if I must actually make this assumption in order to be entitled to

conclude that I believe that P, how can my justification for believing that I believe that P fail to originate, at least to some extent, in my having justification to believe the linking assumption?[14]

Another way of making the same point would be to focus on the idea that epistemically immediate knowledge must not be based on evidence. For if I know that I believe that P on the basis of my knowledge that I judge or conclude that P, then there is a sense in which my knowledge of my own belief is based on evidence. The reason is this: my judging that P is neither identical with nor entails that I believe that P. However, my judging that P normally leads (in the case in which I don't already believe that P) to my forming the belief that P, so the fact that I judge that P raises the probability that I believe that P. It makes it likely that I believe that P and is, in this sense, a reliable sign that I believe that P. But this is just what it is for one thing to be evidence for another.[15] Furthermore, it is not just that my judging that P <u>is</u> evidence that I believe that P. It is also evidence I <u>have</u>, to the extent that I know that I judge that P and am aware, <u>via</u> the linking assumption, of the connection between what I judge to be the case and what I believe. Finally, my knowledge that I believe that P is <u>based</u> on evidence in my possession in the following sense: I know I believe that P <u>because</u> I know that I judge that P, and would not know in this case that I believe that P if I did not know that I have just judged or concluded that P in response to the reasons in favour of P. When this is the basis on which I know that I believe that P my knowledge is not epistemically immediate since it is based on evidence.

<u>3. The Immediacy Intuition</u>

The discussion so far suggests that the transparency procedure, at least as Moran conceives of it, fails to secure either the epistemic or the psychological immediacy of self-

knowledge. It is true that someone who uses this procedure to arrive at knowledge of his own belief doesn't come to know what he believes on the basis of behavioural evidence but there is much more to the notion of immediate knowledge than that. Suppose that this criticism of Moran's proposal is correct. One reaction would be to see it as pointing to the need for an account of self-knowledge that simply does better at respecting the immediacy intuition. However, it would be worth pausing to consider whether this intuition is as robust as Moran and others suppose. The question, in other words, is this: is it even true that our knowledge of our own beliefs and other propositional attitudes is usually immediate? Consider the following claims, which are often treated as equivalent:

(A) We normally know what we think without needing or appealing to evidence.

(B) We normally know what we believe without needing or appealing to evidence.

A possibility that now needs to be considered is that (i) these claims are not equivalent, (ii) the first of these claims is more plausible than the second, and (iii) the plausibility of (A) is mistakenly viewed as lending plausibility to (B). On this account, we should only accept the immediacy intuition if it expresses a commitment to (A). We should reject it, or at least be more cautious about accepting it, if it expresses a commitment to (B).

In one sense, thinking that P is the same as judging that P and is therefore a mental action.[16] One thinks that P in judging that P, and the question whether one's knowledge that one thinks that P is immediate is the question whether one's knowledge that one judges that P is immediate. How, then, does one know that one judges that P? Since judging that P is a mental action, it would be natural to suppose that what enables one to know that one judges that P is a distinctive form of first-personal action-awareness. As Peacocke remarks, 'awareness of your mental actions, such as your awareness that you are deciding, that you

are calculating, and the like, is not merely an awareness that something is happening. It is an awareness that you are doing something, an awareness of agency from the inside' (2007: 364). If I know that I judge that P on the basis of action-awareness of judging that P, my knowledge is not inferential and it is not knowledge based on evidence. I do not infer that I judge that P and my justification for believing that I judge that P does not come from my having justification to believe other, supporting propositions Rather than being a relaible sign that I judge that P, action-awareness of judging that P makes it manifest to me that I judge that P, and the resulting knowledge is both epistemically and psychologically immediate.[17]

Since judging is not believing it does not follow from this that one's knowledge that one <u>believes</u> that P is epistemically and psychologically immediate. Belief is a mental state rather than a mental action, and first-personal action awareness does not supply one with immediate knowledge of one's beliefs. Indeed, it is hard to see how knowledge of one's own beliefs <u>could</u> be immediate. Belief is a form of acceptance.[18] To believe that P is to accept that P, that is, to regard P as true. Yet assuming that P and supposing that P are also modes of accepting that P so what distinguishes belief from other modes of acceptance? As Shah and Velleman argue, part of the answer to this question is that there is a distinctive way in which beliefs are regulated, that is, formed, revised and extinguished: in forming and retaining a belief, 'one responds to evidence and reasoning in ways that are designed to be truth-conducive' (2005: 498). So belief is regulated for truth in a way that other modes of acceptance are not, and being regulated for truth is a broadly dispositional property of beliefs: 'the belief that P tends to be formed in response to evidence of P's truth, to be reinforced by additional evidence of it, and to be extinguished by evidence against it' (2005:

500). Belief-formation can also be influenced by non-rational factors such as wishful thinking, prejudice and phobias but a mental state that is not at least to some extent regulated for truth is not a belief.[19]

This account of belief does not require one to say that beliefs are dispositions, if the point of this characterization is to suggest that beliefs are reducible to dispositions or, even worse, to behavioural dispositions. The proposal that reference to cognitive dispositions is necessary to distinguish belief from other modes of acceptance only requires one to view the relevant dispositions as ones which beliefs have rather than as ones that beliefs are. This proposal also leaves it open whether being regulated for truth is not just necessary but also sufficient for an attitude to count as the attitude of belief. What matters for present purposes is that whether a given mental state is the state of believing that P is at least partly a matter of what dispositions the state has. This proposal helps us to understand what goes wrong in sticking scenarios: judging that P does not always lead one to accept that P, and hence to believe that P, because it is not guaranteed to result in the acquisition of an attitude with the relevant dispositions. The acquisition of these dispositions is effectively blocked by non-rational factors.

One epistemological consequence of this account is that one is not always in a position to know whether one believes that P since one is not always in a position to know that one is in a mental state with the relevant dispositions.[20] It is also unclear, on the present account, how one's knowledge that one believes that P could be immediate. I only believe that P if I am in a mental state which is regulated for truth but how can I know, other than on the basis of evidence, that I am in such a state? How, for example, can I know without any evidence, not only that I accept that P but also that my acceptance of P is such that it

would be extinguished by evidence against the truth of P? In fact, matters are even more complicated than this. Even if I find myself continuing to accept that P in the face of what I recognize as evidence against P it still doesn't follow that I don't really believe that P. Even genuine beliefs can be influenced by non-rational factors, and this makes them even harder to detect.

To sum up, the present suggestion is that belief is a broadly dispositional mental state and can therefore only be detected on the basis of evidence, even when the belief in question is one own. On this account, the immediacy intuition should be rejected, and only seems plausible given a failure to attend to the distinction between believing that P and judging that P. Is this suggestion acceptable? There are several things to be said at this point: the first is that the immediacy intuition is sustained by more than a simple confusion over the relationship between mental states and mental actions. The immediacy intuition is just that, an intuition. In its most vivid form, it is the intuition that my knowledge that, say, I believe my name is Quassim Cassam is both psychologically and epistemically immediate; the idea that believing is a mental action doesn't come into it. Perhaps less of our self-knowledge is like this than is commonly supposed but it is difficult to deny that some of it is both psychologically and epistemically immediate.

If it is possible to know at least some of one's own beliefs immediately then either belief is not a dispositional state after all or it is wrong to suppose that there is any reason in principle why knowledge of one's dispositional mental states could not be immediate.[21] Since there are powerful arguments in support of a dispositional conception of belief, the second of these alternatives is more attractive than the first. The challenge is therefore to explain how, despite the dispositional character of belief, it is nevertheless possible to know

at least some of our own beliefs immediately. It is one thing to think that the extent of immediate self-knowledge is exaggerated, perhaps as a result of a failure to attach sufficient weight to the distinction between belief and judgement, but it would be a mistake to draw the conclusion that knowledge of our own beliefs is never immediate. So the question remains: given what it is plausible to think about the nature of belief how is immediate self-knowledge possible, to the extent (perhaps limited) to which it is possible?

4. Immediacy Explained

Let us return to the problem of antecedent belief. The question is whether I already believe that P, and let us suppose that we are considering a case in which I know, straight off, without any conscious reasoning or inference that the answer to this question is 'yes'. So the case is one in which my knowledge of my own belief appears to be psychologically and perhaps also epistemically immediate. How is this to be explained? Here is one possibility: when I come to believe that P the representation that P enters my belief store or, as is sometimes said, my Belief Box.[22] In order to accommodate the dispositional nature of belief we can stipulate that one's Belief Box is only open to mental representations or states that are responsive to evidence in the specific manner in which beliefs, as distinct from other attitudes, are responsive to evidence.

When I am asked whether I (already) believe that P, what this question calls for is a search of my Belief Box. If the belief that P is found in my Belief Box this leads to the formation of the second-order belief that I believe that P, and this second-order belief might itself end up in my Belief Box. The three questions that now arise are the following:

(a) Who or what is responsible for the search of my Belief Box?

(b) What are the circumstances in which I count as knowing and not merely believing that I believe that P?

(c) How does any of this help to explain either the psychological or the epistemic immediacy of my knowledge that I believe that P?

With regard to (a), the proposal is that the search of my Belief Box is not carried out by me, the subject, but one of my sub-personal monitoring mechanisms, 'a distinct mechanism that is specialized for detecting one's own mental states' (Nichols and Stich 2003: 163). The speed and ease with which this mechanism operates explains the speed and ease with which the second order belief is formed. The discovery of the belief that P in my Belief Box leads directly to the formation of the second-order belief, and there is no mediation by judgement or anything else in the formation of the second-order belief. Since this account draws on aspects of what Nichols and Stich call Monitoring Mechanism (MM) theory of self-awareness I will refer to it as the Monitoring Mechanism account of immediate self-knowledge.[23]

With regard to (b), for my second-order belief to count as knowledge it must satisfy the conditions for knowing. Even if one is sceptical about the prospects for a fully reductive and analysis of the concept of knowledge it is still plausible that there are non-trivial necessary conditions for knowledge and that these conditions include a safety condition: I count as knowing that I believe that P only if my belief that I believe that P could not easily have been false. Since it is easy to conceive of the monitoring mechanism as giving rise to second-order beliefs that satisfy this and other relevant conditions on knowledge it is easy to conceive of my second-order belief that I believe that P as amounting to knowledge that I

believe that P. In this account of self-knowledge all the work is done by one's assumptions about the safety or reliability of the monitoring mechanism.

Turning to (c), the sense in which, according to the MM account, my knowledge that I believe that P is psychologically immediate is straightforward: insofar as my second-order belief results from the operation of a sub-personal monitoring mechanism it is formed without any conscious reasoning or inference. As to the issue of epistemic immediacy, we saw that Moran's account only delivers inferential self-knowledge because it represents my justification for believing that I believe that P as coming from my having justification to believe at least one other proposition, namely the linking assumption. The sub-personal monitoring account certainly does not do that but only because it seems to say nothing at all about whether I am justified in believing that I believe that P. Is the suggestion that there is no justification condition on knowledge, or that there is a justification condition and that what the monitoring account delivers is non-inferential justification?

It is a familiar point that the justification condition on knowledge can be understood either in internalist or externalist terms. The monitoring mechanism account need not deny that knowledge requires epistemic justification, as long as the relevant form of justification is externalist rather than internalist: that is, for one to be epistemically justified in believing that one believes that P it is sufficient that one's second-order belief meets the appropriate safety or reliability condition on knowledge. There is no need for one to <u>believe</u> that one's belief is safe in order for it to <u>be</u> safe so one's justification in this case is not inferential, and there is no reason to think that one's self-knowledge is anything other than epistemically immediate. Externalism about knowledge and justification may strike internalists as a high price to pay for immediate self-knowledge, but the lesson of the discussion so far may well

be that an externalist epistemology is much better placed to account for the immediacy of self-knowledge than internalist alternatives.

We now have a relatively straightforward answer to the question: 'how is immediate self-knowledge possible?'. It is possible to the extent that one's beliefs about one's own beliefs are produced by a reliable Monitoring Mechanism. A creature endowed with such a mechanism would be able to know its own beliefs in a way that is both epistemologically and psychologically immediate, and it is at least arguable that we are such creatures. So the suggestion is not merely that we now have an account of how we could, in principle, come to have immediate knowledge of our own attitudes but also an account of how we actually come to have immediate self-knowledge.

The most pressing of the many questions raised by this account of immediate self-knowledge is whether it can or should accommodate Moran's Transparency Condition. I have represented the MM account as responding to the problem of antecedent belief but, as noted above, there is also the phenomenon of coming to know whether one believes that P by considering the reasons in favour of P. It is one thing to dispute the immediacy of the self-knowledge that is made available by the transparency procedure and another to deny that it is possible for us to acquire self-knowledge in the way that Moran describes. I take it that transparency is a genuine phenomenon, and one that any account of self-knowledge should therefore seek to accommodate.

The problem for the MM theorist is that the procedure that Moran describes is one that takes place at the personal rather than the sub-personal level. It is for me rather than my sub-personal Monitoring Mechanisms to consider the reasons in favour of P, and it hardly needs saying that such mechanisms do not make assumptions about the extent to which our

beliefs are determined by reflection on the reasons in favour of them. Talk of mechanisms for detecting one's own beliefs implies that the beliefs in question are already there, but there is also the case in which I do not already believe that P. Instead, I come to believe that P in the course of considering the question whether I have this belief. In this case, there is no pre-existing belief for my Monitoring Mechanism to latch onto, and the MM theorist does not appear to have a story to tell about this pathway to self-knowledge.

At this point, there are two directions in which the discussion can go. The first would be to try to demonstrate that the MM account can make space for something like the Transparency Condition. The idea would be that when the Monitoring Mechanism searches a Belief Box it is somehow sensitive to the specific ways in which the contents of the box are responsive to evidence. On the other hand, perhaps it is simply a mistake to look for a unified theory of self-knowledge, one that seeks to identify a single way in which we come to know our own beliefs. The alternative would be to accept a hybrid theory on which the MM tells part of the story about self-knowledge, the part that is best equipped to solve the problem of antecedent belief. This leaves room for the idea that avowing a belief is a also way of coming to know that one has it. The fact remains, however, that avowal in Moran's sense is not a source of epistemically or psychologically immediate self-knowledge. When it comes to explaining how immediate self-knowledge is possible the MM account is in better shape. If it only accounts for the immediacy of some of our self-knowledge this only serves to confirm the suspicion that far less of our self-knowledge is immediate than is commonly supposed.[24]

REFERENCES

Alston, W. (1983), 'What's Wrong With Immediate Knowledge?', Synthese 55.

Bar-On, D. (2004), Speaking My Mind: Expression and Self-Knowledge (Oxford: Clarendon Press).

Boghossian, P. (1998), 'Content and Self-Knowledge', in P. Ludlow & N. Martin (eds.) Externalism and Self-Knowledge (Stanford: CSLI Publications).

Boyle, M. (2009), 'Two Kinds of Self-Knowledge', Philosophy and Phenomenological Research, LXXVIII.

Cassam, Q. (2007), The Possibility of Knowledge (Oxford: Oxford University Press).

Davidson, D. (1994), 'Knowing One's Own Mind', in Q. Cassam (ed.) Self-Knowledge (Oxford: Oxford University Press).

Gertler, B. (forthcoming), 'Self-Knowledge and the Transparency of Belief'.

Kelly, T. (2006), 'Evidence', in E. N. Zalta (ed.) The Stanford Encyclopedia, URL: http://plato.stanford.edu/entries/evidence/.

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

Moran, R. (2003), 'Responses to O'Brien and Shoemaker', European Journal of Philosophy, 11.

Moran, R. (2004), 'Replies to Heal, Reginster, Wilson, and Lear', Philosophy and Phenomenological Research, LXIX.

Nichols, S & Stich, S. (2003), Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding of Other Minds (Oxford: Oxford University Press).

Peacocke, C. (1998), 'Conscious Attitudes, Attention, and Self-Knowledge', in C. Wright, B. Smith and C. Macdonald (eds.) Knowing Our Own Minds (Oxford: Clarendon Press).

Peacocke, C. (2007), 'Mental Action and Self-Awareness (I)', in B. McLuaghlin and J. Cohen (eds.) Contemporary Debates in Philosophy of Mind (Oxford: Blackwell Publishing Ltd.).

Pryor, J. (2005), 'There is Immediate Justification', in M. Steup and E. Sosa (eds.) Contemporary Debates in Epistemology (Oxford: Blackwell Publishing Ltd.).

Rosenberg, J. (2002), 'Immediate Knowledge: The New Dialectic of Givenness', in Thinking about Knowing (Oxford: Oxford University Press).

Shah, N and Velleman, D. (2005), 'Doxastic Deliberation', The Philosophical Review 114.

Shoemaker, S. (2003), 'Moran on Self-Knowledge', European Journal of Philosophy, 11.

Williamson, T. (2000), Knowledge and its Limits (Oxford: Oxford University Press).

[1] See, for example, Davidson 1994 and Moran 2001.

[2] This is Moran's conception of immediacy. See Moran 2001: 91. For a different though related account of immediate knowledge see Alston 1983.

[3] This is an example of a how-possible question. See Cassam 2007 for further discussion of such questions.

[4] My knowledge that I am here might be an example of immediate knowledge of a contingent truth. See Boghosssian 1998 for an account of what differentiates this example from immediate knowledge of one's own beliefs.

[5] Shah and Velleman's own response to the problem of antecedent belief is worth quoting: 'If the question is whether I already believe that P, one can assay the relevant state of mind by posing the question whether P, and seeing what one is spontaneously inclined to answer. In this procedure, the question whether P serves as a stimulus applied to oneself for the empirical purpose of eliciting a response. One comes to know what one already thinks by seeing what one says - that is, says in response to the question whether P' (2005: 506). This procedure requires one to refrain from reasoning as to whether P since that reasoning might alter the state of mind one is trying to get at. In addition, testing one's spontaneous response to the question whether P 'may yield good evidence as to whether one already believes that P, but that evidence isn't conclusive: one's first thought upon entertaining a question may be misleading as to one's pre-existing attitude' (2005: 507). On this account, knowledge of what one already believes is clearly based on evidence and therefore not immediate. See Boyle 2009 for some pertinent criticisms of Shah and Velleman's proposal.

[6] There is an illuminating discussion of this phenomenon in Gertler, forthcoming.

[7] As Shah and Velleman observe, 'ordinarily, the reasoning that is meant to issue or not issue in a belief is meant to do so by first issuing or not issuing in a judgement' (2005: 503).

[8] This piece of self-knowledge has yet to be accounted for.

[9] This is presumably not Moran's view. He thinks that knowledge based on observation is not immediate. Yet it is far from obvious that observational or perceptual knowledge is inferential. Part of the problem here is that Moran tends to assume that knowledge that is 'based on' observation is knowledge that is inferred from observational evidence. But this is not the sense in which ordinary perceptual knowledge is based on observation.

[10] This account of non-inferential justification is essentially the one given in Pryor 2005. Pryor correctly assumes that immediate justification is non-inferential justification, and that immediate knowledge is non-inferential knowledge. Moran requires, in addition, that immediate knowledge not be observational. Yet observational knowledge seems to be the paradigm of immediate knowledge, as long as one does not think that all perception involves inference.

[11] As Jay Rosenberg observes, this notion of immediacy 'concerns the de facto origins of bits of knowledge' (2002: 101).

[12] I take it that, on Moran's account, my justification for believing the linking assumption is some form of a priori justifiaction.

[13] Dorit Bar-On argues that the transparency method is 'epistemically rather indirect' to the extent that it implies that self-judgements 'are arrived at on the basis of consideration of wordly items' (2004: 113). In my terms, the indirectness implied by this characterization of the transparency method is primarily psychological. On my account the transparency method is epistemically indirect, but not quite for the reason given by Bar-On.

[14] Shoemaker raises a similar question about the role of the linking assumption in Moran's discussion. See Shoemaker 2003: 401.

[15] As Williamson remarks, what is required for e to be evidence for the hypothesis h is that 'e should speak in favour of h' and should itself have 'some kind of creditable standing' (2000: 186). In probabilistic terms, e speaks in favour of h if it raises the probability of h. Kelly points out that 'the notion of evidence is that of something which serves as a reliable sign, symptom, or mark of that which it is evidence of' (Kelly 2006).

[16] This is of course not to suggest that all thinking is judging. For example, one can think about P without judging that P.

[17] Even if the argument of this paragraph is correct does it not leave open the possibility that my knowledge of what I judge is perceptual and, in this sense, not immediate? There are two things to be said about this. The first is that it is not obvious that perceptual knowledge should be classified as mediate knowledge, and even if there is a legitimate sense in which perceptual knowledge is not immediate it is very different from the sense in which inferential knowledge is not immediate. The second thing to say is that it is arguable, in any case, that action-awareness is not a form of perceptual awareness. For more on this see Peacocke 2007.

[18] Here and in the next few paragraphs I follow Shah and Velleman 2005. The view that belief is a form of acceptance is also endorsed by Peacocke. He correctly remarks, however, that 'belief' can also be used for a feeling of conviction. See Peacocke 1998: 72 n.5.

[19] Shah and Velleman also draw attention to the influence on belief of what they call 'evidentially insensitive processes' (2005: 500). Their example of such a process is wishful thinking.

[20] In Williamson's terminology (which is different from Moran's) 'transparency' is the thesis that 'for every mental state S, whenever one is suitably alert and conceptually sophisticated, one is in a position to know whether one is in S' (2000: 24). He goes on to argue that transparency fails for the state of believing since 'the difference between believing P and merely fancying P depends in part on one's dispositions to practical reasoning and action manifested only in counterfactual circumstances' (2000: 24). In effect, Williamson's point is that the dispositional dimension of believing makes trouble for what he calls transparency. My point is that it makes trouble for immediacy.

[21] There is a more general issue here about whether, when F is a dispositional feature of objects, it is possible to know immediately that a particular object is F. For example, on a dispositional view of colour, something's being red consists in its being disposed to look red in normal conditions. It is certainly possible to see, without any conscious reasoning or inference, that a ripe tomato is red, and one's knowledge in this case is psychologically immediate. Epistemic immediacy is trickier. Perhaps my belief that the tomato is red depends for its justification on my being justified in believing that conditions are normal. This would make my knowledge epistemically mediated even though it is based on seeing that something is the case. Some have concluded all perceptual knowledge is inferential since one always relies on the assumption that conditions are normal. No such assumption is required for self-knowledge.

[22] A Belief Box represents a functionally characterized set of mental states. As Nichols and Stich point out, positing such a box 'does not commit a theorist to the claim that .... the states are spatially localized in the brain, any more than drawing a box in a flow chart for a computer programme commits one to the claim that the operation that the box represents is spatially localized in the computer' (2003: 11). Thanks to Tim Williamson for the suggestion that the Belief Box account can accommodate the dispositional character of belief.

[23] As Stich and Nichols point out, a good theory of self-awareness needs to be able to explain the fact that 'when normal adults believe that P, they can quickly and accurately form the belief I believe that P' (2003: 160). In order to implement this ability, 'all that is required is that there be a Monitoring Mechanism (MM) that, when activated, takes the representation P in the Belief Box as input and the representation I believe that P as output' (2003: 160-1). The Monitoring Mechanism simply has to copy representations from the Belief Box and embed copies of them in a schema of the form 'I believe that...'. Stich and Nichols do not draw attention to the consequences of their view for the issue of immediacy, but it seems obvious that if a Monitoring Mechanism is sufficiently reliable to produce knowledge of one's own beliefs then the knowledge to which it gives rise is both psychologically and epistemically immediate. It was Timothy Williamson who first drew my attention to the possibility of exploiting something like the MM theory to explain the immediacy of self-knowledge. He does not, however, endorse the present approach to self-knowledge.

[24] For helpful comments I thank Bill Brewer, Steve Butterfill, Tim Crane, Naomi Eilan, Christoph Hoerl, Hemdat Lerman, Guy Longworth, Johannes Roessler and Matthew Soteriou. Thanks also to Tim Williamson for some very helpful discussions of this topic.