Review Essay:
SELF-KNOWLEDGE AND ITS LIMITS

Quassim Cassam, *Self-Knowledge for Humans* (Oxford: Oxford University Press, 2014), 272 pages. ISBN: 9780199657575 (hbk.). Hardback: £ 30.00.

John Doris, *Talking to Our Selves* (Oxford: Oxford University Press, 2015), 240 pages. ISBN: 9780199570393 (hbk.). Hardback: £ 30.00.

In recent decades there has been a surge of interest in self-knowledge among philosophers of a broadly 'analytic' orientation. This interest was driven at first by questions about whether privileged access to the contents of one's mental states was compatible with 'externalist' views of mental content, and later by a number of influential books and papers offering more systematic accounts of self-knowledge in general, chief among these the works brought together in Richard Moran's *Authority and Estrangement*. Moran tries in that book to reorient the philosophical discussion of self-knowledge away from narrowly epistemological concerns about the nature and scope of privileged access, arguing that self-knowledge should also be a main concern of philosophical moral psychology, and that our conception of the epistemic status of self-knowledge will be incomplete as long as it fails to account for the place of self-knowledge in a person's social and psychological well-being.

An important theme in Moran's book is that what he calls the 'fundamental asymmetries' between first- and third-personal perspectives on a person's thoughts, attitudes, actions, feelings, etc. – for example, the way that a person usually knows 'without observation' what she believes or is intentionally doing, that the self-ascription of thoughts and actions is usually somehow authoritative, and so on – are neither exceptionless and neatly definable nor always a ground of first-personal privilege. That is, Moran's position is meant not just to allow for, but also to explain the inevitability of, situations in which a person may need such things as observation, inference, or the testimony of others to know her true thoughts or feelings, and also to explore why there are some things a person is in a position to know better about others than she does about herself, so that self-ascriptions of these sorts tend to be outright disprivileged in contrast to their third-personal counterparts.

The works under review, both of which aim to deconstruct excessive and unrealistic philosophical conceptions of first-personal privilege and replace these with a picture of self-knowledge that emphasizes the frequent opacity of one's own psychology, both overlook or ignore the similar themes in Moran's work, despite identifying that work as representative of the view of first-personal authority they wish to critique. Though both authors prefer to address themselves to 'families' of views or general 'tendencies' in philosophical work rather than engaging the details of how various philosophers have developed these ideas, in Quassim Cassam's book the identification of Moran as a target is quite explicit: Moran is the chief representative of what Cassam calls 'Rationalist' conceptions of self-knowledge which, Cassam claims, have distorted its nature by overstating the privilege of our

self-knowledge at the expense of understanding its real psychological significance. If this account were correct, then Moran's project would be a failure on its own terms.

In fact, the construal of Moran as a pure 'Rationalist', like Cassam's reconstruction of the 'Rationalist' approach in general, is an inaccurate caricature. Cassam's 'Rationalist' tries to account for self-knowledge on the assumption that the average human being is of the species *homo philosophicus*, a 'model epistemic citizen' who can know her own beliefs because the following two conditions are satisfied in her case (see SKH, p. 4):

(i) What HP believes is what she ought rationally to believe, what she wants is what she rationally ought to want, what she fears is what she rationally ought to fear, and so on.

(ii) HP knows or justifiably believes that what she wants is what she rationally ought to want, what she fears is what she rationally ought to fear, and so on.

The 'Rationalist' makes these assumptions in order to explain the *transparency* of self-knowledge, or how the self-ascription of beliefs, desires, and so on is made not on introspective grounds, but rather by considering the 'external' facts in reference to which these attitudes are supposed to be rationalized. Given (i) and (ii), *homo philosophicus* could come to know what she believes by determining what she rationally ought to believe, as Gareth Evans suggests in a memorable passage from *The Varieties of Reference*:

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 procedure I have for answering the question whether P …[1]

For the 'Rationalist', answering Evans's outwardly directed question is a way for a person to determine what she rationally ought to believe. Knowing this, and given the assumptions (i) and (ii) above, *homo philosophicus* can know her beliefs in the sort of way that Evans describes.

Of course, human beings are not *homines philosophici*, and so 'Rationalism' thus construed is hopeless as an account of human self-knowledge. Cassam documents our irrationality extensively by appeal to the experimental literature on heuristics and biases, but that humans have – to put it mildly – a tendency to reason badly is not a fact that we needed laboratories to discover. What *would* be surprising would be if careful philosophers had built their theories on the premise that humans' beliefs, desires, fears, etc. are always or even generally guided by reason. And Moran, in any case, makes no such assumption. Rather, the most Moran assumes is that many of a person's attitudes will conform to what she *takes to be* the balance of the reasons for and against them – an assumption that may be false, but isn't challenged at all by showing that we often believe things for bad reasons, as long as in those cases we mistake those bad reasons for good ones. Based as it is in this inaccurate account of Moran's position, Cassam's main argument against it fails.

---

[1] G. Evans, *The Varieties of Reference* (Oxford: Oxford University Press, 1982), at p. 225.

What is more valuable in Cassam's book than this misplaced critique of Moran's supposed 'Rationalism' is his invitation to emphasize what he calls 'substantial' self-knowledge in contrast to the 'trivial' self-knowledge that he thinks philosophers have recently been too focused on. 'Trivial' and 'substantial' here have to do with the moral weight or personal importance of the attitudes a person is supposed to have knowledge of: an example of the former would be the knowledge that I believe it is raining, or want to have pizza for lunch; of the latter, the knowledge that I am happy in my marriage, or harbor racist or sexist attitudes that I would explicitly disavow. Cassam argues, and here he is correct, that much of our substantial self-knowledge can't be accounted for in the way suggested in the quote above from Evans. For a person to determine e.g. how satisfied she is with her career she must do more than just consider whether that career is satisfying, since she might judge that her career is satisfying when she reflects on it in this explicit and self-conscious way while remaining dissatisfied in a way that doesn't reflect her deliberate assessment.

So Cassam is right that a philosophical account of substantial self-knowledge cannot assume that our attitudes are perfectly responsive even to our (frequently mistaken) assessments of the reasons for and against them: knowing these important truths about our 'real selves' requires something more than our ordinary capacities for self-conscious reasoning. This again is not much of a strike against Moran, however, since as noted above he positively emphasizes those cases where a person's mind is 'opaque' to her in just these ways; e.g.:

> The person who feels anger at the dead parent for abandoning her, or who feels betrayed or deprived of something by another child, may only know of this attitude through the eliciting or interpreting of evidence of various kinds. She might become thoroughly convinced, both from the constructions of the analyst, as well as from her own appreciation of the evidence, that this attitude must indeed be attributed to her. And yet, at the same time, when she reflects on the world-directed question itself, whether she has indeed been betrayed by this person, she may find that the answer is no or can't be settled one way or the other. So, transparency fails because she cannot learn of this attitude of hers by reflection on the object of that attitude. She can only learn of it in a fully theoretical manner, taking an empirical stance toward herself as a particular psychological subject.[2]

In the context of passages like this one, Moran's extended discussion in *Authority and Estrangement* of the 'trivial' self-knowledge that e.g. I believe it is raining is intended not as an account of what human self-knowledge is like all the time or even in general, but rather as an illustration of a certain *paradigmatic* form of self-knowledge that is grounded in self-conscious reflection rather than introspection, inference, self-observation, or the advice of an analyst. This knowledge is trivial but the philosophical discussion of it is not, any more than the philosophy of perception is 'trivial' when it tries to explain how a person can know by sight that a nearby object in good lighting is a tea-cup, even as there are so many more important matters that perception cannot so easily reveal.

---

[2] R. Moran, *Authority and Estrangement: An Essay on Self-Knowledge* (Princeton: Princeton University Press, 2001), at p. 85.

The fact (if it is a fact) that transparent self-knowledge is paradigmatic, and perhaps also fundamental in being a form of self-knowledge without which a person might not count as a rational agent, and without which other forms of self-knowledge might not be possible at all, of course does not show that those other forms of self-knowledge do not deserve serious philosophical consideration. And so it is fair to criticize philosophers for focusing too much on explaining transparency and not enough on understanding cases that are not so paradigmatic, but are nevertheless of immense personal and philosophical importance. The best thing about Cassam's book is the invitation it offers to consider those cases more fully, and to see them as no less important to an account of self-knowledge than the cases where such knowledge is easier to come by. The shame is that he so grossly misrepresents the position he is using as a foil.

John Doris's target in *Talking to Our Selves* is not rationalism but a position he calls *reflectivism*, according to which it is a condition of morally responsible agency that a person have self-knowledge of her reasons for acting. Doris's reflectivism is, once again, not a doctrine that is held in all its details by any particular philosopher, but unlike Cassam Doris outlines a view that corresponds to some real tendencies in serious philosophical thought. Still, according to Doris the 'traditional conceptions of practical rationality' that assume reflectivism are unable to account for the ways human beings actually order our lives, guided as we often are by 'an unconscious at odds with deliberate intention' (TOS, pp. 16, 5). Doris argues that the reflectivist will have to regard unconsciously motivated behaviors as non-agential, and thus that reflectivism leads to skepticism about the existence of responsible agency. The solution, he claims, lies in a nonreflectivist position on which 'the expression of values associated with the exercise of agency need not be a reflective process' (TOS, p. 33) – a position he goes on to develop in the second part of his book.

My focus here will be on the negative part of Doris's argument. (I see no reason why the reflectivist cannot take on board many of Doris's nonreflectivist ideas, though as a supplement to a sensible reflectivism rather than an outright replacement for it.) The strategy there is to demonstrate the prevalence of what Doris calls *self-ignorance*, then argue that any view on which 'the exercise of human agency requires accurate reflection' (TOS, p. 19) – requires, that is, an accurate self-knowledge of one's reasons for acting, or of the values that an action expresses – will treat self-ignorance as undermining agency. According to Doris, to avoid the skeptical threat that emerges when we take self-ignorance seriously, we must reject those reflectivist theories in favor of some nonreflectivist alternative.

How, though, is this skeptical threat supposed to arise? Doris claims that the possibility of widespread self-ignorance is supported by studies in experimental social psychology that reveal 'influences on behavior that are both unconscious and unexpected … from the perspective of practical reasoning' (TOS, p. 43). These influences, he claims, are 'causes of behavior that are not plausibly taken as reasons for behavior', since a person would not 'be willing to treat [such a] consideration as a justification for their judgment or behavior' (ibid.). Moreover, many cases of self-ignorance are *practically relevant*, in the sense that were a person to become aware of such an unconscious influence 'she would behave – or argue, judge, feel, etc. –

differently' than she will when she is unaware of the influence (TOS, p. 21). The self-ignorant agent is not self-directed in the way the reflectivist supposes that agency requires: her thinking and behavior are controlled instead by things outside her ken.

Spelled out in more detail, Doris's argument is that the reflectivist should be troubled by cases where the outputs of unconscious cognitive processes are *incongruent* with those of self-conscious reasoning, and where these unconscious outputs *bypass* deliberation so that 'behavior is influenced by a process that the actor is unaware of, and would not recognize as a reason justifying the behavior, were she so aware' (TOS, p. 51). According to Doris, the reflectivist must treat such a case of ignorance with bypassing as a *defeater* of the exercise of agency, since 'the causes of [the actor's] cognition or behavior would not be recognized by the actor as reasons for that cognition or behavior, were she aware of those causes at the time of performance', and so the reflectivist will hold that 'the exercise of agency does not obtain' in such a case. More generally: 'If the presence of defeaters cannot be confidently ruled out for a particular behavior, it is not justified to attribute to the actor an exercise of agency. If there is general difficulty in ruling out defeaters, skepticism about agency ensues' (TOS, pp. 64-65).

Whether the reflectivist must regard self-ignorance as incompatible with agency will depend, however, on the extent of that ignorance, as well as the degree of accuracy in reflective self-knowledge that the reflectivist takes agency to require. For example: in the Stroop task, a subject's report of certain features of a perceptual stimulus, such as its color, will be influenced by task-irrelevant features of the stimulus, such as which word it is – a subject will e.g. report the red color of the word 'red' more quickly than the red color of the word 'blue'. We might note further that this ignorance can be practically relevant, say if the goal is to have one's response time be determined just by the actual color of the stimulus: in such a situation, a subject who knew of the unconscious influences would adopt a different strategy, e.g. of deliberately ignoring task-irrelevant features. But the reflectivist should not therefore think that the self-ignorant subject's response is not an instance of agency, or even that the subject's agency is significantly compromised, unless the influence of the task-irrelevant feature is quite large. How large? That depends, again, on the details of the case and the reflectivist position in question. But the point is that not just any unconscious influence will be enough to threaten agency, even according to a rather austere reflectivism. Nor will identifying just a few agency-undermining influences be enough to support a skeptical threat to the existence of agency *in general*, which according to Doris is supposed to happen here.

As another illustration of this point, consider the findings Doris discusses about how intuitively irrelevant factors like a person's name can influence their life choices, e.g. of a career or a place to move to (see TOS, pp. 54-56 and 71-73). Supposing for the sake of argument that the relevant influences are unconscious, and also that they work contrary to explicit deliberative processes rather than complementing them by, say, helping to break 'ties' or making certain possibilities more salient, then these will be cases of incongruence with bypassing. But whether a reflectivist should count these as defeaters to the exercise of deliberative agency, as opposed to influences that make an agent somewhat irrational or limit the scope of her agency without altogether eliminating it, will depend on the extent to which a

person's choices are driven by unconscious influences *rather than* self-conscious thought. And in the cases at issue, the causal role assigned to unconscious factors in determining a person's choices is very small: Doris cites a finding that women were 18% more likely than they would be by chance to move to states with names resembling their first names, and 36% more likely for states that were perfect matches;[3] but still it appears that about 94% of the women in that study's dataset moved where they did for reasons unrelated to the 'name effect'.[4] (Note that this dataset included *only* women named Florence, Georgia, Louise, or Virginia who moved to Florida, Georgia, Louisiana, or Virginia. That last percentage would have been even higher if it also concerned women with those names who moved to states other than these four.) As long as a lot of that other variance in these choices can be explained by conscious reasoning, the reflectivist should not be troubled by these results.

Doris is aware of this difficulty. He does not, however, address it head-on by detailing any wide range of actual cases where unconscious influences *do* defeat agency totally rather than just partially – that is, cases where the reflectivist should accept that agency is altogether 'swamped' (see TOS, p. 68), and not just subject to some extra-agential influences that work against the agent's self-conscious thought. Instead, Doris argues that for all we know there *could* be cases like that, where the unconscious and incongruent factors that bypass conscious deliberation also overrule its force, and that the burden is on the reflectivist to rule this possibility out. Thus he writes that his skeptic about agency 'needn't deny the existence of partial defeaters having only limited impact on agency, she need only to insist on the difficulty of establishing, in any particular case, that it is only such comparatively benign influences which obtain' (ibid.). And again, now in more detail (I have added the bracketed numbers for ease of reference):

> Once we see that [1] there are some arbitrary influences on cognition and behavior, we are bound to admit that [2] there may be others; if something like *that* can make a difference, there could be *many* goofy influences in any particular instance. While the impact of each goofy influence may be statistically small, just as with medical interventions, [3] the aggregate effect may be quite potent; for all one knows, any decision may be infested by any number of rationally and ethically arbitrary influences. (TOS, p. 64)

This skeptical argument has some questionable company. Consider: Once we see that [1'] there are some cases where sensory perception is inaccurate, we are bound to admit that [2'] there may be others; if the senses can mislead us in a case like *that*, there could be *many* other cases where they do the same. While the impact of each

---

[3] B.W. Pelham, M.C. Mirenberg, and J.T. Jones, "Why Susie Sells Seashells by the Seashore: Implicit Egotism and Major Life Decisions", *Journal of Personality and Social Psychology*, 82 (4), 469-487.

[4] That is: of the 33,412 women in the dataset, 12,991 moved to states whose names resembled their own, but 10,999 of these would have been expected to move to those name-resembling states anyway; thus name resemblance made a difference in an estimated 6% of cases. (Again, the unrepresentative dataset explains why the overall proportion of women moving to name-resembling states is so high.) For these numbers, see Pelham et al., Table 6 (p. 475). Thanks to Sam Sims for this analysis.

sensory illusion may be small, [3'] the aggregate effect may be quite potent; for all one knows, any sensory experience may be infested by any number of illusory factors. This train of reasoning appears cogent, but even Descartes' skeptical meditator denied that it warranted any general mistrust of the senses: noticing the particular cases where the senses mislead us, such as when things 'are hardly perceptible, or very far away', should not lead us to question the 'many others to be met with as to which we cannot reasonably have any doubt', at least not just on the basis of these everyday sensory inaccuracies. Yet Doris means for his skepticism to be different from the fantastical 'evil demon'-based stuff that the Cartesian skeptic ultimately settles on: his skeptical hypothesis is, he writes, 'not a loopy (and perhaps massively unlikely) proposition like skeptical propositions involving Demons, Matrices, or envatted brains. Rather, it is a "live" hypothesis … vetted by the relevant experts, and judged by a substantial number of them to be about as likely as competing hypotheses' (TOS, p. 66).

But *what* has been expertly vetted, and judged likely, is not the likely widespread existence of cases where the aggregate effects of unconscious processing swamp those of self-conscious deliberation, but only the (probably) widespread existence of unconscious influences on thought and behavior which, though philosophically very interesting (as are cases of perceptual inaccuracy), do not individually come anywhere close to canceling the force of deliberation. To conclude that since we know from experiment that these phenomena are likely prevalent, therefore they *could* be everywhere, and moreover occur in such concentration and with such unidirectional force that we are never agents at all, would be like claiming that the sensory inaccuracies revealed by the science of perception *could* be the rule rather than the exception, and aggregate in such a way that we never truly perceive the world around us. The latter argument can be met in a Moorean fashion: the skeptical hypothesis is loopy, not live, because I can see quite clearly that this is a hand. And the reflectivist can respond to Doris's argument in the same way: in a wide range of everyday cases it is quite clear that we are agents, and that our self-conscious deliberation guides our thought and behavior. That our behavior in these cases may also be influenced by something other than deliberation, and our perception not always perfectly accurate, does not even tend to show that we are not usually self-knowing agents and world-knowing perceivers.

The point of this criticism is not to deny that extra-agential influences on our choices can have a cumulative impact that makes them morally, psychologically, and politically very important. The question is where that cumulative impact lies. For example, in a recent paper the social psychologists Anthony Greenwald, Mahzarin Banaji, and Brian Nosek address criticisms of the power of the 'Implicit Association Task' (IAT), a widely used measure of implicit stereotypes, to predict discriminatory behavior.[5] Their reply emphasizes that it is 'system-level' patterns of discrimination, not individual choices, of which the IAT is a useful predictor: while using this measure to predict biased choices by a given person would 'risk undesirably high

[5] A.G. Greenwald, M.R. Banaji, and B.A. Nosek, "Statistically Small Effects of the Implicit Association Test Can Have Societally Large Effects", *Journal of Personality and Social Psychology*, 108 (4), 553-561, at pp. 557-558.

rates of erroneous classifications', these worries 'diminish substantially as sample size increases', e.g. when the concern is with correlations that emerge through the aggregate decisions of larger populations of similarly biased individuals. Thus one may accept the reality of implicit bias, and agree with Doris that its effects are 'practically and theoretically large' (TOS, p. 63), while still holding that our specific decisions of e.g. whom to hire, arrest, convict, etc. are largely responsive to reason-giving factors that are present to consciousness, and thus that even our biased choices will count as expressions of agency according to the reflectivist standard.

What is especially frustrating here is that the nature and extent of our self-ignorance, and the importance of this for theorizing about agency and practical reasoning, could all be discussed without the problematic skeptical frame that Doris gives it. Consider again the phenomenon of implicit racial prejudice (see TOS, pp. 56-58). That e.g. decisions about whom to hire appear to be affected by how 'African-American-sounding' a candidate's name is, reveals how things can go terribly wrong in deliberation about an important decision, with significant costs to society when such effects cluster around widespread norms that are reinforced by discriminatory social structures. But these effects do *not* suggest that any individual person whose hiring decision is influenced by implicit prejudice is therefore not an agent in that decision, or that this decision does not largely reflect her self-conscious assessment of the candidates' qualifications. (As with the 'name effect', most of the variance in hiring is determined by the evaluator's assessment of a candidate, with unconscious bias tipping the scales in cases where those qualifications are indecisive.) What they *do* suggest is that the influence of prejudice can impair our rationality, and make our decisions worse than they could be, with the potential for significant injustice to individual persons and larger social groups. That is something for philosophers to take very seriously even if it really is not, as Doris suggests, 'surprising from the perspective of practical rationality and ethical theory' (TOC, p. 59) – for after all, even paradigmatically rationalist and reflectivist philosophers like Plato, Augustine, Descartes, and Kant all emphasized how often reason fails to act as our guide. It can be fair to criticize philosophers with noble conceptions of reason's proper role in thought and action for failing to grapple sufficiently with these very important phenomena. But that grappling should not have to take the form of fending off an imaginary skeptical threat.

What Doris describes in his opening chapter as the fragmentation of the *psyche* that was the story of 20th-century psychology, like the demonstration of heuristic irrationalities that Cassam discusses, is not really a new story at all: versions of it were told before in Plato's account of the divided soul, Augustine's confession of weakness and self-deception, and the 'lassitude' of Descartes' meditator as he lapsed from his philosophical conclusions back into the opinions of ordinary life. That story is told as well in the works of the great poets, playwrights, and novelists, Austen and Eliot chief among them, who all saw how often the things we say to ourselves are just attempts to obscure our true motives and thereby see ourselves as noble. As Cassam and Doris emphasize, these tendencies are most common, and their effects most powerful, when the objects of our reflection are characteristics of ourselves that we regard as ethically weighty: this makes it especially challenging for us to manage in these value-laden self-assessments to see

ourselves as we really are. (Once again, this is a main theme in Moran's work too.) What social psychology does is operationalize these tendencies, giving us a way to measure them under controlled conditions. If this helps empirically minded philosophers to take the unconscious more seriously, and steer us away from sanitized and simplistic depictions of a self transparent to itself, philosophical theorizing will benefit from this development. It is good to have these books invite us to that task.[6]

John Schwenkler
Department of Philosophy, Florida State University
jschwenkler@fsu.edu

[6] For helpful feedback and discussion, thanks especially to Michael Bishop, Brian Boeninger, Michael Brownstein, Nick Byrd, Gabriel De Marco, Kenny Easwaran, Craig French, Kyle Fritz, Jeff Haines, Sophie Horowitz, Bryce Huebner, Daniel Kirchner, Mark Lance, Neil Levy, Edouard Machery, Eric Marcus, Casey O'Callaghan, Laurie Paul, Geoff Pynn, Sam Sims, Nicholas Sparks, Jay Spitzley, Matthew Taylor, and Marshall Thompson.