2014/2015 | 136TH SESSION | VOLUME CXV

PROCEEDINGS OF THE ARISTOTELIAN SOCIETY

Edited by MATTHEW SOTERIOU (WARWICK)

ISSUE NO. 3

Self-Knowledge, Agency and Self-Authorship

SACHA GOLOB (KCL)

DRAFT PAPER

AS

WWW.ARISTOTELIANSOCIETY.ORG.UK

MEMBERSHIPS | LATEST ISSUES | PODCASTS | VIRTUAL ISSUES | THE JOINT SESSION

PROCEEDINGS OF THE ARISTOTELIAN SOCIETY
136TH SESSION

ISSUE NO. 3
VOLUME CXV
2014 / 2015

SELF-KNOWLEDGE, AGENCY AND SELF-AUTHORSHIP

SACHA GOLOB
KING'S COLLEGE LONDON

MONDAY, 18 MAY 2015

17.30 - 19.15

ROOM 349
SENATE HOUSE
UNIVERSITY OF LONDON
MALET STREET
LONDON WC1E 7HU
UNITED KINGDOM

This event is catered, free of charge, & open to the general public

CONTACT
mail@aristoteliansociety.org.uk
www.aristoteliansociety.org.uk

BIOGRAPHY

Sacha Golob is a Lecturer in Philosophy at King's College London; prior to that he was a Research Fellow at Peterhouse, Cambridge. His research focuses on the intersection between the history of philosophy and contemporary philosophy of mind, action and ethics. He is the author of *Heidegger on Concepts, Freedom and Normativity* (CUP 2014), and the editor of the forthcoming *Cambridge History of Moral Philosophy* (CUP 2016)'.

EDITORIAL NOTE

SELF-KNOWLEDGE, AGENCY AND SELF-AUTHORSHIP

SACHA GOLOB

> In short, it is a matter of placing the imperative to "know oneself" – which to us appears so characteristic of our civilization – back in the much broader interrogation that serves as its explicit or implicit context: What should one do with oneself? What work should be carried out on the self? (Foucault, 'Subjectivity and Truth')

THIS PAPER ADDRESSES the question of a subject's knowledge of his or her own mental states. My interest, in particular, is in an appeal to the concepts of mode and activity when explaining our ability to self-ascribe beliefs. Ultimately, I argue for an agency account of self-knowledge that avoids the excessive rationalism of positions such as Moran's and Boyle's. Before getting underway, some restrictions on scope. My discussion deals solely with propositional attitudes; I say nothing about sensations. The main reason is that the contemporary agency accounts in which I am interested typically impose a similar restriction: as we will see, this is because the notion of activity on which they rely is intimately linked to a responsivity to reasons which sensations lack. Indeed, the natural tactic for such theorists is to follow Kant in arguing for two distinct stories about self-knowledge: one, to be examined here, concerning "consciousness of what the human being does", the other, suitable for sensations, "consciousness of what he undergoes" (Kant 2006, p.161).[1]

I. THREE RESPONSES TO EVANS ON TRANSPARENCY

I want to approach the debate via Evans' famous example:

> 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 in 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. (Evans 1982, p.225)

[1] As Boyle puts it, such theories should deny the "uniformity assumption", the assumption that all forms of self-knowledge will be amenable to the same type of explanation (Boyle 2009, p.141).

Call this 'the Evans case' or EC. Like many important philosophical claims, Evans' point can seem simultaneously obvious and incredible. It can appear obvious because I take it that his description is accurate: if asked such a question, I would proceed as he says. As Bar-On notes, this is not unusual:

> Asked whether I find my neighbour annoying, I would ponder her actions and render a verdict….In general, in addressing questions about what I think, believe, want, prefer, feel, and so on, I concern myself not with me and my states, but rather with the world outside myself. (Bar-On 2004, p.11)

Yet, on reflection, this is puzzling for two reasons. First, there is the problem of self-ascription. In EC, I arrive at a verdict on whether I believe that *P* by establishing whether *P*. But there are countless cases where *P* holds and yet I don't believe it: the mere truth of *P* is neither inductively nor deductively linked to my endorsing it. Another way to put the worry is this: I have addressed a question about one issue, my own mental states, by looking at a completely different issue, geopolitics. O'Brien aptly dubs this the "two topics problem": how, from content about the world, have I arrived at a conclusion about the self? (O'Brien 2007, p.103). Second, there is a problem as to how judgment meshes with belief. Suppose one thinks of judgments as conscious acts or processes and beliefs as standing dispositional states.[2] The worry then arises: could not someone judge that *P* and yet this judgment fail to be sufficiently 'internalised' to yield a belief that *P*? Hence Peacocke:

> 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. (Peacocke 1998, p.90).

In this section, I will discuss a particular line response to these two challenges.

I want to introduce the response in which I am interested by distinguishing it from two more extreme options. The two extreme options are characterised by their stance on what, following Byrne, I'll call "neutrality": an account is neutral iff it explains self-knowledge using premises which are not themselves specified in terms of the subject's awareness of his or her mental states (Byrne 2005, p.94). The first of the two extremes flatly rejects neutrality. Consider this from Brentano:

[2] For example, Cassam 2011a.

> The fact that the mentally active subject has himself as object of secondary reference, regardless of what else he refers to as his primary objects, is of great importance. As a result of this fact there are no statements about primary objects which do not include several assertions. If I say, for example, "God exists" I am at the same time attesting to the fact that I judge that God exists.(Brentano 1973, p.215)

Following Kant, phenomenological writers often frame discussions of content in terms of "objects" (for example, Kant 1998, A55/B79). At least in this passage then, Brentano's suggestion is that the move in cases such as EC is not from *P* to *I believe that P*; rather it is from *I judge that P* to *I believe that P*. This approach faces numerous problems. First, as many authors have stressed, our judgments typically seem transparent in the Moorean sense – their content is solely world-directed. As Sartre puts it, "I am plunged into the world of objects…there no place for *me* on this level" (Sartre 1972, p.49). Second, it is clear that no such account can explain self-ascriptive content; it rather concerns itself entirely with the transition from a tacit to an explicit awareness of such. Third, the account is essentially spectatorial: in addition to considering Russian tanks, I necessarily also have another object before my eyes, my own acts. Of course, these are only a "secondary object", something that entails for Brentano complex epistemic restriction. I cannot focus attention on my own acts in the way I can the Russian tanks, for example: they are, so to speak, confined to the periphery of my vision (Brentano 1973, p.215). Nevertheless, the view remains vulnerable to the charge, pressed by authors like Moran against introspectivist theories, of:

> [A]n essentially superficial view of the differences between my relation to myself and my possible relation to others.(Moran 2001, p.91.

Self-knowledge is construed as privileged perception; the only difference between myself and a perfect mind reader who could watch my mental states unfolding before his eyes is that no such other perceiver exists.

The second of the two extremes, in contrast, enthusiastically embraces neutrality: in other words, it explains EC as indeed progressing legitimately from *P* to *I believe that P*. I have in mind Byrne's view on which such transitions are "strongly self-verifying" since "inference from a premise entails belief in that premise" (Byrne 2011, p.206). I cannot do justice here to the ingenuity of Byrne's position. Instead, I want simply to indicate my agreement with Boyle and O'Brien that it nevertheless violates a key desideratum: our account of EC should explain not only why the transition is safe, but why the subject might perceive the move as a rational one, i.e. as resting on an "intelligible relation" between premise and conclusion (Boyle 2011b, p.231; O'Brien 2005, p.591).Of course, the inference im-

mediately becomes intelligible if the premise is not simply *P*, but the fact that *I* accept that *P*; but then we are back to something like the Brentanian position.

I have discussed two responses to EC. I want now to introduce an attractive, if elusive, compromise: perhaps the move is neither simply from *P*, nor from some content which already contains a self-ascription. Rather, the premise is *P* – but presented under a certain mode or from a certain standpoint. As noted, the phenomenological tradition often frames claims about content in terms of objects. By extension, the broad view I am considering here is often expressed by saying that self-awareness is not a form of object-awareness. For example, Sartre:

> [T]his consciousness of consciousness…is not *positional,* which is to say that consciousness is not for itself its own object. Its object is by nature outside of it. (Sartre 1972, p.41)

Husserl makes similar remarks, as does Heidegger when outlining his own account of experience as "*selbstweltlich*" (for example, Heidegger 1994, p.96). Strikingly, this tactic is also prominent among contemporary analytic authors. O'Brien, for example, suggests that it "is something about the mode – in contrast to content – of the state or activity" that is the key to handling EC (O'Brien 2007, p.126). The trick, as Boyle observes, is that such approaches complicate the status of neutrality – whilst the first order state remains solely world-directed in terms of its content, the reference to its mode of presentation is not "genuinely non-committal as to the nature of the subject's mental states" (Boyle 2011b, p.233). In short, the move in EC is not simply from *P*, but from *P* under some specific mode of presentation, and it is this which renders it intelligible.

How might this be cashed? One immediate option is to develop the idea of a non-positional or non-objective form of awareness. In recent work, Boyle draws on Sartre to make the following proposal:

> [H]er concluding that [there will be a third world war] must involve an implicit awareness of her taking this answer to be correct. For if she were not aware of this…then the question would still remain open for her, and her deliberation would not have concluded. So although *what* she represents as the case is a proposition about the non-mental world, her manner of *representing* it depends on an implicit awareness of her own determination about what is correct. (Boyle 2014, p.23)

Boyle's "reflectivist" view is that to reach the fully fledged self-ascription that *I believe that P* the subject needs simply to reflect on this prior, non-objectual awareness of her own orientation vis-à-vis *P*. Boyle's proposal is an extremely interesting one, but I am sceptical as to whether it can work.

First, is there really a sufficient explanatory gap between being aware that *I believe that P* and being aware that I take my deliberation on *P* to be settled to avoid assuming what is to be explained? Second, how does such non-positional awareness relate to more familiar propositional content? Does it have accuracy conditions, and if so what is the story regarding error with respect to it? What factors prevent it from being reducible to tacit propositional content, a reduction which would again bring it uncomfortably close to assuming what it seeks to explain, namely fully fledged self-ascription? Consider the difficulties in establishing that perception is non-propositional or non-conceptual even when one can draw on all the distinctive features of visual awareness. Of course, sound answers might be available to these questions. But I want instead to try to develop another way of making good the 'mode of presentation' approach. This is the task of section II.

II. MORAN AND THE AGENCY THEORY: A QUALIFIED DEFENCE

My aim in this section is to sketch Moran's position, and to defend it against some familiar objections. In section III, I will then press a different, and I think more persuasive, line of attack.

Whilst Moran does not specifically rely on 'mode of presentation' terminology, his view is nevertheless well classified an instance of the compromise strategy of section I. This is because he holds that EC is not simply a transition from *P* to *I believe that P*, but rather from *P* addressed from within a "practical, deliberative" perspective to *I believe that P* (Moran 2001, p. xvii). More specifically, he argues that the normal method, in both the statistical and evaluative senses of that term, by which we arrive at knowledge of our own propositional mental states is not a matter of discovering "some antecedent fact about oneself", but rather one of "making up" our mind (Moran, 2001, p.58). In short, I learn whether I believe that *P* by considering the reasons for or against *P* and reaching a verdict on them: insofar as this verdict determines my belief, I can know of the latter by establishing the former (Moran 2003, p.405). The self-directed question is *transparent* to the world-directed one: "I get myself in 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" (Evans 1982, p.225). First person authority is thus not that of a privileged spectator, but of an agent: in judging that *P*, I make it the case that *I believe that P* (Moran 2001,p.124). Of course, not all such beliefs are so reached: Sophia might learn of her beliefs about her father through therapy in which she takes "an empirical stance on herself as a particular psychological subject" among others (Moran 2001, p.85). But such cases

are normal neither in the statistical sense nor the evaluative one: after all, if Sophia cannot arrive at those beliefs directly by reflecting on the facts about her father, it suggests that the beliefs are not fully rational (Moran 2001, p.108).

Moran's proposal has attracted enormous interest, and many objections. I think that some can be dealt with quickly. For example, the key to the proposal is that in determining what I take to the case, I determine what I believe. As Moran stresses, this does not imply a doxastic voluntarism: there may, when I consider the facts, be only one thing to think and thus to believe (Moran 2011, p.3). Furthermore, the position equally holds if there is only one thing to think, and I simply see that without deliberation in any extended sense: one might talk more neutrally not of 'judging' but of 'taking something to be the case', of an immediate response to the world and the reasons it provides. For example, Shoemaker raises the following worry:

> I know and believe that I believe that I am wearing pants…But it is hard to think of circumstances, other than those of a dream, in which it could be a question for me whether I believe this. I would also have a hard time saying what reasons I have for believing it. And I cannot think of any good sense in which it is 'up to me' whether I believe. (Shoemaker 2003, p.396).

As I see it, Moran can simply reply that I take it to be the case that I am so dressed, i.e. I take in the world and thereby regard this claim as both compelling and obvious. In short, the visibility of the conceptual space within which I conclude that *P*, whether clear or murky, makes no difference to the proposal.

More troublesome, in contrast, are examples in which I know that *I believe that P*, where this belief is not readily groupable with cases such as therapy, and yet where my knowledge of it is not a function of an explanatorily prior verdict that *P*. One class of such cases involve reason responsive and world-directed judgments where I have nevertheless not arrived at them *by* responding to reasons, and thus where my capacity to self-ascribe the relevant states cannot be a function of such a response. Cassam gives the example of the thought that today is the first of the month which simply pops into my head as I write; this is reason responsive in that it would be extinguished by countervailing evidence, but my awareness of the belief is not a function of looking at the world and making a call, even an immediate one (Cassam 2011b, p.5). Boyle cites a firm belief in some historical fact where I can no longer remember any grounds for it: instead, my only basis for judging that *P* is my prior awareness that *I believe that P*, so inverting Moran's 'world to self' order

of explanation (Boyle 2014, pp.6-7). A second class of cases seek to force one to read see the self-knowledge problem in what Moran calls a "theoretical" sense, i.e. as requiring the report of "some antecedent fact about oneself" rather than a decision on the world-directed question of whether *P* (Moran, 2001, p.58). For example, Shah and Velleman suggest that Evans' original example is ambiguous: the questioner might have meant "do I already believe that *P* (i.e. antecedently to considering this question)" (Shah and Velleman 2005, p.16). If this is the case, then my now reaching a verdict on *P*, in line with Moran's proposal, is actually prohibited since "that reasoning might alter the state of mind that one is trying to assay" (Shah and Velleman 2005, p.16). Reed offers a related challenge, but one where the pressure to treat the inquiry "theoretically" comes not from the way the original question is framed, but from a weakening in the agent's epistemic standing. In Reed's example, Penny has written a book in which she defends views on some complex topic, including whether *P*. Later, she is asked what she believes about *P*; whilst "she knows she has staked out a position with respect to it…[she] simply cannot recall it now" (Reed 2010, p.176).[3] Suppose, further, that Penny returns to her office and looks in her book, and sees there the verdict that *P*. The default position, Reed claims, is that this is still Penny's belief: not only does she thus learn of her belief without directly considering whether *P* (instead she looked in a book, as she might equally have done when seeking to learn about some other agent), but this is actually the rationally virtuous path to self-knowledge in such a case (after all, she has forgotten many of the intricacies of the debate and is not well-placed now to address the world-orientated issue directly) (Reed 2010, p.178).

What might be said in defence of Moran here? With respect to Cassam's calendar example and other 'out of the blue thoughts', the best strategy is a divide and conquer one. Either such thoughts are ways of taking the world to be, i.e. actions and commitments based on a consideration of reasons even if that consideration is involuntary and done at a glance, or they can be treated as purely passive, "as merely entertainings of content that…come before the mind – as perceptions or memory images might" (O'Brien 2013, p.96). If the former, they are susceptible to Moran's account: whilst I have not gone through any explicit deliberation, I still take it to be the case that *P* and I can then self-ascribe this belief in line with the transparency procedure. If the latter, they can be accommodated by whatever additional account is needed to treat phenomena such as sensations. One option for dealing with Boyle's case, meanwhile, is to argue that I originally arrived at self-knowledge via the transparency procedure. For example, I came to know I believed that *P* by judging that *P* on the

[3] Reed 2010: 176.

basis of testimony or other evidence, even though I am now able only to recall the outcome, not the evidence. Moranian transparency would thus remain the explanatory primary mechanism for self-knowledge – it is just that here, we have memory only of its outputs, not its workings.

What of the second group of counter examples? There the key is whether it is coherent to self-ascribe something as my belief without simultaneously taking a stance on the question of its plausibility. As Boyle puts it: "I do not recall what I believe about whether P unless I recall what now looks to me to be the truth as to whether P" (Boyle 2011a, p.10).[4] This principle seems plausible, at least in the current context (I say more on this in section III below). To see its impact, consider again the Reed case.[5] If the statements in the book are to be Penny's beliefs, they must now look like the truth to her; in other words, we cannot see her as learning about her *beliefs* unless she simultaneously takes the claims in the book to be accurate. But once that is conceded, it supports a reading of the story on which her book is a source of evidence which Penny is using to now take a view on *P*, and so learn of her own belief via the standard Moranian model. Reed objects that were this the case, we should expect Penny to check not or not just her own book when she returns to her office, but one by the "acknowledged master of the field"; he or she would, after all, be the best source of evidential testimony (Reed 2010, p.177). Yet there seem good ground why Penny might still privilege her own book even if she is treating it as evidence for a decision as to whether *P*. There is a social, rational and habitualised pressure towards consistency and so absent significant evidence of error, we typically give extra weight to views we held earlier.[6] Furthermore, even if Penny regards Jane as the "acknowledged master of the field", the data in her own book is precisely that which she has previously felt to be most persuasive: given the likely psychological and epistemic continuity between her earlier and present selves it make sense to start there (in effect, she is taking testimony within a framework which she knows she finds plausible).

[4] There is no tension with Boyle's own attack on Moran via the 'forgotten history case'. As Boyle presents it there, I do indeed recall "what now looks to me to be the truth as to whether P": the worry is that such conviction is explained by, rather than explanative of, my knowledge that I believe that *P*.

[5] For parallel discussion of Shah and Vellemans's point in the light of this principle, see Moran 2011, pp.223-4.

[6] If there were significant evidence of error, for example if Penny had doubts about her earlier competence or if the master's book was published later and had access to a broader data set, it seems that she would consult it in preference to her own.

III. BROADENING THE ROLE OF AGENCY: INFERENCE AND INTERNALISATION

In section II, I considered various counter examples to Moran's position. I want now to return to the two core challenges regarding EC which I set out in section I. In both cases, one can see how Moran might face a problem. First, one might deny that Moran adequately explains self-ascription: suppose I deliberate and conclude that *P*. How does this yield the conclusion *I believe that P*? Where *exactly* does the self-directed content enter?[7] Second, one might deny that Moran adequately explains the judgment/belief relationship. Suppose I judge that *P*; what has Moran said against the possibility that this would fall short of my believing that *P*? Space prohibits treatment of both of these issues, and so in this final section I will focus on the second, and assume that some compatible treatment of the first can be found. Ultimately, I will sketch a position that retains Moran's focus on agency whilst diluting the rationalistic tenor of his position.

I want to approach the issue via question of inference. Prima facie, self-knowledge seems immediate in some distinctive sense: it is clear, for example, that there is typically no extrapolation from external behaviour. Yet, as Cassam observes, matters are not so simple.

> My knowledge that *P* is epistemically immediate only if my justification for believing that *P* does not come, even in part, from my having justification to believe other, supporting, propositions. My knowledge that *P* is psychologically immediate only if it is not acquired by conscious reasoning or inference. If I come to know that I believe that *P* by employing the transparency procedure then my knowledge does not appear to be immediate in either sense. (Cassam 2011b, p.3).

I think that Cassam is half right here. I think that Moran can reasonably insist on psychological immediacy: not only might I take *P* to be the case at a glance, the fact that there are cases where I deliberate about whether *P* does not suffice to show that the transparency transition itself, from *P* to *I believe that P*, relies on any conscious reasoning. Instead, it will typically be an automatic move, since, for reasons I will come to in a second, the assumption that we are entitled to it is deeply enmeshed with the nature of deliberation. But with respect to epistemic immediacy, I think Cassam is correct: Moran's position is not epistemically immediate because the subject needs to be aware of the principle that when she judges *P* it follows that she believes that *P* (Cassam 2011b, p.12). Now, I agree with Moran that such a commitment is a necessary condition on being said to undertake an act of deliberation or judgment, as opposed to just making some facial gestures or noises – hence my remark just made about automaticity

[7] Both Byrne 2011, p.203 and Shoemaker 2003, p.398 raise this worry.

and psychological immediacy. But even if the link between judgment and belief is a "Transcendental assumption of Rational Thought", it remains a proposition which the subject must endorse, and thus renders the transparency process non-immediate in at least Cassam's epistemic sense.[8]

I do not regard this as problematic in itself: nothing just said conflicts with the indisputable datum that self-knowledge involves no inference from external behaviour, and there is a clear story as to why arriving at self-knowledge in Moran's way might still be psychologically immediate. The resultant position also remains better off, I would suggest, than accounts such as Boyle's discussed in section II: there is no need, for example to cash some kind of non-positional content. Instead, there is ordinary world-directed content, and an automatic inference from it to self-ascription. However, once it is conceded that the transparency procedure relies on what Cassam calls a "linking assumption" about the capacity of judgments to generate beliefs, it seems natural to press the point further: to what degree, might this link fail? Recall, for example, Peacocke's story of prejudice regarding foreign degrees. One option would be to try to downplay such cases: Boyle suggests that we either interpret the person as first both judging and believing that the foreign degrees are as good as domestic ones and then later changing later her mind to judge and believe something else, or as never genuinely judging that they are just as good and so again never exhibiting a judgment/belief misalignment (Boyle 2014, p.19). But this seems unattractive: one can set up the example such that the cosmopolitan judgment and the chauvinistic behaviour are synchronic, and there seems no independently motivated reason to deny that a rational agent who undertakes what is in every other regard an act of judgment is not genuinely doing so simply because of its misalignment with his or her beliefs.

Moving beyond individual cases to the structural issue, I agree with Cassam that the most compelling reason for postulating a potential divergence between judgment and belief will be the ontological assumption that, whilst judgments are occurrent mental acts or events, beliefs are to be cashed in terms of dispositional states (Cassam 2011b). Given this taxonomy, it seems immediately plausible both that someone might judge that *P* and yet have a longstanding disposition to act in ways that imply a belief that *not P*, and that even multiple acts of judgment might fail to reconfigure sufficiently sedimented dispositions, particularly when these are embedded in causal and conceptual links to many other affective and representational states – religious beliefs are a natural example. The im-

[8] One might respond that there is no need for the subject to endorse it – simply for it to be the case. But, as noted when discussing Byrne, it seems that this is not enough: we need to explain EC in a way which captures its first person intelligibility.

pact of thinking of beliefs as dispositions will be amplified if one looks not just at propositional contents, but at attitudes too: once those are treated dispositionally, it seems likely that whether a given state is a belief or only a useful fantasy will depend "in part on one's dispositions to practical reasoning and action manifested only in counterfactual circumstances", something over which the fact that we now judge that *P* seems to give us no particular authority (Williamson 2000, p.24). It is worth noting, incidentally, how these type of concerns differ from Reed's argument against Moran, treated in section II. The problem with Reed's example is that the context of the story requires us to understand beliefs as commitments – Penny is trying to establish what she believes so that she can inform her colleague and debate the position with him. In such a context, Boyle is surely right: if Penny is to believe that *P*, *P* must now look to her right – or else her 'belief' might fail to count as a reason, as something she might propose and defend. But when beliefs are glossed as behavioural dispositions, it becomes natural to think that what I take to be the case and what I actually do can diverge.

I want now to suggest a specific way to see the judgment/belief relation given these results, one which preserves what is most attractive in Moran whilst avoiding the excessive rationalism. The proposal is this. To judge that *P* is to exert a distinctive kind of causal power on oneself. Where there are no countervailing causal forces in play, for example strongly networked affective or motor intentional patterns, this power is sufficient for believing that *P*, i.e. for acquiring the requisite dispositions. When Tom judges from the map that Paris is in France he acquires the corresponding belief; whereas when he judges that his keys are now stored upstairs after years of being kept by the door, his beliefs will lag his judgments just as they will when the phobic judges that the plane is safe even as he sits there sweating and shaking. There are, of course, many very deep issues regarding causation and the mental which I cannot discuss here. But one can see how the resultant position is, in an important sense, an agency model of self-knowledge: in the good case, I know *that I believe that P* by making it the case that I do so through my act of judging that *P*. I can endorse, for example, the following, just as Moran can:

> [T]he primary thought gaining expression in the idea of 'first-person authority' may not be that the person himself must always 'know best' what he thinks about something, but rather that it is his business what he thinks about something, that it is up to him. (Moran 2001, p.124).

Of course, there is an inferential component in my account: to make the self-ascriptive move I must assume that my judgments have this power. But, as we have seen, so must Moran himself. In most cases, agents will

in fact move between judging that *P* and self-ascribing the corresponding belief automatically; where they are aware of countervailing forces, they will rightly be hesitant (consider belief ascription by agents who have just been prompted by reading the implicit bias literature). The link will also depend on details about the agent: some individuals may have particularly 'strong wills', i.e. causally efficacious capacities to determine their behaviour through conscious reflection. Finally, as Nietzsche observes, there is also a social dynamic in play: those individuals able to sustain a close alignment of judgment and belief possess the "prerogative to promise", to take on *commitments*, through reasoning, which have cash value at the level of their own behaviour (Nietzsche 1994, 2/2). As McGeer, whose position is probably closest to the one defended, puts it:

> First-person judgements – judgements we make about what to believe or desire – have a certain 'commissive quality': they are judgements made in the indicative mode – I do believe this – that commit us to speak and act in ways commensurate with those judgements. (McGeer 2007, p.87)

One way to put the point is this: a first person judgment is not a prediction as to my behaviour, but an undertaking and attempt to exert a certain kind of control over such. Insofar as this exercise of agency is successful, my judging that *P* is my believing that *P*, and I may self-ascribe the latter state by assuming this link.

Talk of causality in this context may bring to mind the familiar debate surrounding self-blindness. But matters here are different than with classic causal introspectivist theories. Even an agent in whom the judgment/belief link had totally broken down would neither be self-blind (since he might, for example, have privileged first person knowledge of his passive states through whatever mechanism is appealed to to handle sensations), nor totally passive (since he would still be able to judge and respond to reasons at the occurrent level). However, the position does entail that something like a global state of akrasia, intellectual and practical, is metaphysically possible: I do not regard this as a problem, partly since I doubt we have clear intuitions over such a case.

There are, of course, many concerns one might have about such a proposal, and its causal dimension in particular, and I want now to address two recent arguments on the topic.

First, Boyle argues that a causal model of the judgment/belief link renders problematic various facts about the temporality of agency. Boyle's arguments are highly intricate and I cannot deal with each one here, but I want to highlight one central contention he makes. Given that "a cause must precede its effect", the causal model entails that:

> I act on the basis of an (apparent) reason for believing P that I now possess, in a way that will only later result in my believing P. Since it is possible for me to acquire new information, or for my assessment of the grounds for P to change, there need not be any time here at which I reasonably believe P…To appeal to our consistency over time or the small probability that new considerations will present themselves in the time that elapses seems to introduce irrelevant complications into our account of the rationality of doxastic agency. (Boyle 2011a, pp.12-13).

I accept in the vast majority of cases, considerations regarding consistency or probability do seem irrelevant. But this can be explained in several ways. Most obviously, they might seem irrelevant because when the step between judgment and belief is *only* a function of the metaphysical principle that a cause precedes its effect, the resultant gap is simply not a salient one – in other words, considerations regarding probability are irrelevant not in the sense that they have no place here, but rather in the sense that we do not bother to appeal to them since there is no ground to think they alone are sufficient to generate a misalignment. Furthermore, they might seem irrelevant since when we study justification we typically present it as a relation among propositions, bracketing the issue of their temporal realisation outside the 'third realm'. They would thus be irrelevant because we are used to abstracting away from them in order to address other questions.

Second, both Moran and Boyle argue that the exercise of a merely causal power over our beliefs fails to acknowledge the intimacy of the judgment/belief link.

> [T]here is surely an intuitive contrast between my power to govern whether I have a stomach ache and my power to govern whether I believe P: whereas in the former case my control over the relevant condition is at best indirect, in the latter, one wants to say, my control may be direct. (Boyle 2011a, p.17).

Clearly, there are many differences between altering my belief through judgment and altering my digestion through diet. The question is whether it is a necessary condition on accommodating them that one abandons the approach to agency I have suggested. A two pronged response seems attractive here. On the one hand, I can stress the distinctive ways in which judgment modifies beliefs which have no parallel in cases like the stomach ache, and yet which seem fully compatible with my theory: for example, judging that *P* might lead me to acquire a new concept, which might in turn cause the sematic structure of my beliefs to alter, something that is clearly not possible in the case of digestion. On the other hand, I can argue that the gap between the belief case and the digestive one is not as black and white as Boyle suggests. Just as I manipulate the environment

to reduce the likelihood of indigestion, there are countless devices which I employ to bridge the potential gap between judgment and belief. I have in mind here the type of detailed, historical analysis which someone like Foucault offers of different practices of diary keeping, of memory games, of public proclamations and rituals, of mutual agreements to observe and correct – each of these taking on a highly specific and distinctive form in, say, a medieval Christian context, or a Stoic one, or a modern one.

This brings me to a final, broader, point. Moran obviously recognises that judgment might fail to yield the corresponding belief – for example, in cases of akrasia. I can likewise accommodate his point that:

> [I]n the case of ordinary theoretical reasoning, which issues in a belief, there is no further thing the person does in order to acquire the relevant belief once his reason has led him to it. (Moran 2001, pp.118–9)

This is because in the ordinary case, you need do no more than judge; the conditions are such that this will yield, without friction, the requisite belief. In a sense, then, what is at stake is how unusual or defective we consider cases of imperfect judgment/belief alignment. One way to frame the issue is in terms of rationality or psychological health: as Moran sees it, to believe that *P* just when you judge that *P* is "both the normal condition and part of the rational well-being of the person" (Moran 2001, p.108). I think that notions of rationality are too ambiguous here to be much use: if John and Tom both make conceptually incoherent and racist judgments, but Tom's prior training means that he alone cannot in fact bring his beliefs and behaviour into line with them, there is at least some sense in which he is rationally better off . I think the notion of "well-being" is also a very loose one. As McGeer notes, we can imagine both cases in which an ability for seamless judgmental self-governance sustains a pattern of disturbing rationalisations, and cases in which a willingness to see oneself as an empirical object, only partly guided by deliberation and very much prey to other forces, is clearly "psychologically healthy, even admirable" (McGeer 2007, p.92). So I would prefer to frame it like this. The transparency procedure retains a distinctive and central role in the context of self-knowledge. This is because a form of agency, judging that *P*, will typically make it the case that I have the relevant belief, and assuming I know this, I can thus self-ascribe on the basis of world-directed evaluation, exactly as in EC. Yet we should simultaneously recognise that this mode of agency is part of a broader story, one concerning the variety of methods through which individuals seek to author or determine themselves: judgment is only a defeasible device for doing so and it is never found unsupported by those other more external, indirect tools for shaping our belief, such as repetition or ritual, in which, to borrow a phrase from Moran, something

"is inflicted on me, even if I am the one inflicting it" (Moran 2001, p.117). Thus the remark from Foucault which began this paper:

> In short, it is a matter of placing the imperative to "know oneself" – which to us appears so characteristic of our civilization – back in the much broader interrogation that serves as its explicit or implicit context: What should one do with oneself? What work should be carried out on the self? (Foucault 1997, p.87)

Department of Philosophy
King's College London
WC2R 2LS
sacha.golob@kcl.ac.uk

REFERENCES

Bar-On, Dorit 2004: *Speaking My Mind*. Oxford: Clarendon Press.

Boyle, Matthew 2009: 'Two Kinds of Self-Knowledge'. *Philosophy and Phenomenological Research*, LXXVIII, pp.133-63.

--- 2011a: 'Making up Your Mind and the Activity of Reason'. *Philosophers' Imprint*, 11, pp.1-24.

--- 2011b: 'Transparent Self-Knowledge'. *Proceedings of the Aristotelian Society Supplementary Volume*, LXXXV, pp.224-41.

--- 2014: 'Transparency and Reflection', Unpublished MS.

Brentano, Franz 1973: *Psychology from an Empirical Standpoint*. London: Routledge.

Byrne, Alex 2011: 'Transparency, Belief, Intention'. *Proceedings of the Aristotelian Society Supplementary Volume*, LXXXV, pp.202-21.

Cassam, Quassim 2011a: 'Judging, Believing, Thinking'. *Philosophical Issues*, 20, pp.80-95.

--- 2011b: 'Knowing What I Believe'. *Proceedings of the Aristotelian Society*, CXI, pp.1-23.

Evans, Gareth 1982: *The Varieties of Reference*. Oxford: Clarendon Press.

Foucault, Michel 1997: 'Subjectivity and Truth'. In Rabinow, P. (ed.) *The Essential Works of Michel Foucault Vol. 1, Ethics: Subjectivity and Truth*, pp.87-92. London: Allen Lane.

Heidegger, Martin 1994: *Phänomenologische Interpretationen zu Aristoteles*. Frankfurt a.M.: Klostermann.

McGeer, Victoria 2007: 'The Moral Development of First-Person Authority'. *European Journal of Philosophy*, 16, pp.81-108.

Moran, Richard 2001: *Authority and Estrangement*. Princeton University Press.

--- 2003: 'Responses to O'Brien and Shoemaker'. *European Journal of Philosophy*, 11, pp. 402-19.

--- 2011: 'Self-Knowledge, 'Transparency', and the Forms of Activity'. In Smithies and Stoljar, (eds.) *Introspection and Consciousness* , pp.211-236. Oxford University Press.

Kant, Immanel, 1998: *Critique of Pure Reason*, trans. Guyer and Wood. Cambridge University Press.

--- 2006: *Anthropology from a Pragmatic Standpoint*, trans. Louden: Cambridge University Press

Nietzsche, Friedrich 1994: *On the Genealogy of Morality*. Cambridge University Press.

O'Brien, Lucy 2005: 'Self-Knowledge, Agency and Force'. *Philosophy and Phenomenological Research*, LXXI, pp. 580-601.

--- 2007: *Self-Knowing Agents*. Oxford University Press.

--- 2013: 'Obessive Thoughts and Inner Voices'. *Philosophical Issues*, 23, pp.93-108.

Peacocke, Christopher 1998: 'Conscious Attitudes, Attention, and Self-Knowledge'.In Wright, Smith and Macdonald (eds.) *Knowing Our Own Minds*, pp.63-98. Oxford University Press.

Reed, Baron 2010: 'Self-Knowledge and Rationality', *Philosophy and Phenomenological Research*, LXXX, pp.164-81.

Sartre, Jean-Paul 1972: *The Transcendence of the Ego*, trans. Williams and Krkpatrick. New York: Octagon Books.

Shah, Nishi. and Velleman, David 2005: 'Doxastic Deliberation'. *Philosophical Review*, 114, pp.497-534.

Shoemaker, Sydney 2003: 'Moran on Self-Knowledge'. *European Journal of Philosophy*, 11, pp.391-401.

Williamson, Timothy 2000: *Knowledge and Its Limits*. Oxford University Press.

THE ARISTOTELIAN SOCIETY