Sacha Golob (sacha.golob@kcl.ac.uk)
Forthcoming in the Proceedings of the Aristotelian Society
Preprint – Please Cite the Published Version

SELF-KNOWLEDGE, TRANSPARENCY AND SELF-AUTHORSHIP

SACHA GOLOB

Forthcoming in the Proceedings of the Aristotelian Society

> 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')

ABSTRACT

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 sketch an agency account of self-knowledge that avoids the excessive rationalism of positions such as Moran's and Boyle's.

This paper addresses the question of a subject's knowledge of his or her own mental states. My interest is particularly in an appeal to the concepts of mode or activity when explaining our ability to self-ascribe beliefs. Ultimately, I sketch 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 will 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 deny the "uniformity assumption" (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. 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: this concerns the relation between claims about mental states and claims about the world. 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 different issue, geopolitics. O'Brien aptly dubs this the "two topics problem" (O'Brien 2007, p.103). Second, there is a problem as to the relation between different types of mental state and our ability to move from knowledge of one such type to knowledge of another. Suppose one thinks of judgments as conscious acts or processes and beliefs as standing dispositional states. 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*?

> 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).

Applied to EC specifically, the worry is that my simply "answering the question whether p" might not, contra Evans, put me in "a position to answer the question whether I believe that p".

I want to introduce a particular line of response to these two challenges. I will do so by distinguishing it from two more extreme options. The 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:

> 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 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": 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: 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: we should explain not only why the EC 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 immediately 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 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 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 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. But how might this be cashed? Section II considers three answers.

II. Three Versions of the 'Mode of Presentation' Approach

I will now examine three versions of such an approach, beginning with Moran's extremely influential work – given its prominence, I will devote some time to establishing where exactly Moran's view succeeds and where it might fall short.

Whilst Moran does not specifically employ 'mode of presentation' terminology, his theory is nevertheless well classified an instance of the compromise strategy of section I. This is because he reads EC as not simply a transition from *P* to *I believe that P*, but rather from *P* addressed 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, 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 addressing the practical or deliberative question of whether *P* is to be believed, and I do that by considering the reasons for or against *P* and reaching a verdict on them: insofar as this verdict determines my belief, I can know the latter, 'inner' fact, by establishing the former, 'outer' one. (Moran 2003, p.405). The self-directed question thus is *transparent* to the world-directed one: as in EC, "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). 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 subnormal in both senses: after all, if she 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).

Some of the charges standardly raised against Moran 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 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 whether *P*. Years 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).[2] 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 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

[2] Reed 2010: 176.

recall what now looks to me to be the truth as to whether P" (Boyle 2011a, p.10).[3] This principle seems plausible, at least in the current context. To see its impact, consider again the Reed case.[4] 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*: she thus establishes what she believes by looking at the facts in the world, exactly as Moran contends. Reed objects that were that the case, we should expect Penny to check not just her own book, but one by the "acknowledged master of the field"; 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 text even if she is treating it as evidence 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. 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).

I have defended Moran against a number of objections, but there remains a real problem with his account: where *exactly* does the first person content enter? As Byrne puts it: "Suppose that I examine the evidence and conclude that there will be a third world war. Now what?" (Byrne 2011, p.203). The issue is occluded in so far as one starts by asking 'what do I believe about *P*?' and then moves, in line with Moranian transparency, to look at the worldly facts as to *P*. Set up in this direction, the reference to the self is already embedded in the initial question. But a full account of propositional self-knowledge surely requires that we can also move in the other direction; namely, explaining how, from purely world-orientated judgments, we might arrive at self-ascription. What Moran lacks, in other words, is a good gloss on the 'I': how do I get from the world-orientated verdict that *P* or even that *P* is to be believed to claims about myself?

This brings me to the second proposal I want to discuss, one expressly intended to make good this deficiency. In recent work, Boyle draws on Sartre to defend the idea of a non-objectual awareness of the self as implicated in even apparently world-directed attitudes.

> [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*, namely her taking it to be settled. Boyle's proposal is an extremely interesting one, but I remain unconvinced. 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

[3] 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*.
[4] For parallel discussion of Shah and Velleman, see Moran 2011, pp.223-4.

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.

The third and final version of the compromise strategy shares Boyle's emphasis on simply making explicit what was tacitly present, but rather than awareness of one's own doxastic orientation, it appeals to activity and control. The basic idea is that when judging that *P*, the subject is aware of his or her own activity, and that it is this awareness which provides the basis for self-ascription. "Basis" here can be read as 'independent warrant', as in O'Brien.

> There is a form of awareness had by creatures capable of controlling their actions, mental and physical, that is independent of any capacity of the creature to understand the term or concept 'I', that is both non-conceptual and non-perceptual in nature and yet that is capable of immediately warranting the self-ascription of the action that the creature is aware of in this way. (O'Brien 2007 p.76)

"Basis" can also be read along more Kantian terms as a constitutive relation: the undertaking of a certain activity, rational judgment, just is the forming of those connections, such as norms of consistency and coherence, which define the subject. As Kitcher puts it, "it is through trying to make sense of sensory data that cognizers come to combine representations and so to create the relations across their states that are the hallmarks of single subjects" (Kitcher 2011, p.262). But the obvious question for both variants is how exactly we understand the idea of such action awareness. It must be sufficiently thin both that it is plausibly present even in engaged or undeliberated acts (or else we would be unable to self-ascribe beliefs based on such judgments), and that it avoids collapsing back into the Brentanian content approach canvassed in section I. A further, underlying, danger concerns ambiguities in the key concept of activity. To give a single example, the Kantian definition on which "we are active when our mental life displays sensitivity to reasons" (Raz 1997, p.218), is neither obviously necessary nor obviously sufficient for action in the sense of self-initiated behaviour: it is not obviously necessary since animals may act in the latter sense whilst plausibly lacking at least our sensitivity to reasons, and it is not obviously sufficient due to cases where, as McDowell famously put it, we are "saddled" with perceptual content that is nevertheless normatively structured.

Ultimately, of course, the key will lie with a closer analysis of what counts as content or object-awareness. It seems plausible that to be aware of options *x*, *y*, and *z* as open to control and manipulation is to be aware of those very options as showing up in a certain way, rather than to be aware of them in conjunction with any fourth thing; yet the more one stresses the difference from the Brentanian content approach of section I, the greater the risk there will not be enough left to warrant the move to the self. I am optimistic, although not completely convinced, that some requisite notion of action awareness can be found. This would provide a way of handling the shift from *P* to *I judge that P*. So we now have at least a map indicating how to proceed. But rather than follow the details of that path through, I want to keep the discussion at the level of our overall orientation towards Evans' problem. Specifically, I want to say something about the second aspect of EC, the move from

knowledge of one class of mental states to knowledge of another class, and how a simultaneous appeal to another, thicker notion of activity might also speak to that.

III. Internalisation and Self-Authorship

Recall Peacocke's story of prejudice regarding foreign degrees. Someone considers the evidence and honestly judges that *P*; but they cannot be said to know that they believe that *P* since their behaviour makes it plain that in fact they do not. 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 a compelling reason for postulating a potential divergence between judgment and belief is the fact 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 impact 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* gives us no particular authority (Williamson 2000, p.24). It is worth noting, incidentally, how these types of concern 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. 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. 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 remarks from Moran:

> [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).

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 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, given a viable account of self-reference such as that canvassed above, I can self-ascribe the latter state by assuming this link. The resulting combination might be called a 'causal constitutivism'. Talk of causality in this context may bring to mind the familiar debate surrounding self-blindness. But matters here are different than with classic 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 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 doubt we have clear enough intuitions over such a case for this to be problematic.

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 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 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. Someone like 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 thus, given something like the account canvassed in section II, that 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). Hence 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)

Foucault's own point is, unsurprisingly, independent of the issues treated in section II. But one can see how they might mesh well together. On the one hand, one would accept a thin notion of action, that which provides the basis for self-ascription even in the sensory deprivation tank of the Anscombian literature. On the other, one would also acknowledge a thicker notion of action, a way of recognising the many forms of self-authorship within which that thin awareness might be manifest, and on which the self is not a given, but a project of stylisation and control.

Department of Philosophy
King's College London
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*, pp.87-92. London: Allen Lane.

Heidegger, Martin 1994: *Phänomenologische Interpretationen zu Aristoteles.* Frankfurt: 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*. Cambridge University Press.

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

Kitcher, Patricia 2011: *Kant's Thinker*. Oxford 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.

Raz, Joseph 1997: 'When We Are Ourselves'. *Proceedings of the Aristotelian Society*, 71, pp.211–227.

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

Sartre, Jean-Paul 1972: *The Transcendence of the Ego*. 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.