Habituation and First-Person Authority

Jonathan Webber

in

Time and the Philosophy of Action

edited by Roman Altshuler and Michael Sigrist

Routledge, 2015

Abstract

Richard Moran's theory of first-person authority as the agential authority to make up one's own mind rests on a form of mind-body dualism that does not allow for habituation as part of normal psychological functioning. We have good intuitive and empirical reason to accept that habituation is central to the normal functioning of desire. There is some empirical support for the idea that habituation plays a parallel role in belief. In particular, at least one form of implicit bias seems better understood as a case of habituated belief than as a mere association or an example of what Tamar Gendler calls 'alief'. If there is to be genuine first-person epistemic authority over persisting mental states, therefore, an alternative account to Moran's is required in the case of desire and perhaps in the case of belief. More generally, the neglect of habituation in recent philosophy of mind is a symptom of the need for philosophers to take the temporal structure of rational agency more seriously.

How can you gain knowledge of the contents of your own mind? Are any ways of doing so available only to you? If so, are claims about your mental states arrived at in this way more authoritative than claims made on other grounds? These traditional questions of self-knowledge and first-person authority have been reinvigorated over the past few decades by the idea of the transparency of belief: you can answer some questions of whether you believe that p by considering whether p is true; but nobody else can discover whether you believe that p simply by considering whether p is true. What does this observation about the attribution of beliefs show us about self-knowledge and about the mind in general?

My central claim in this paper is that the leading recent account of transparency and first-person authority rests on a form of mind-body dualism that should be rejected. Richard Moran has argued that the first-person authority afforded by transparency can be understood only if we think of ourselves primarily as agents. It is not in our capacity as knowers that we find our most significant access to the contents of our own minds, but in our capacity as doers. Our epistemic authority about our own mental states is derivative of our agential authority over them. This is a significant departure from the philosophical framework that has dominated discussions of self-knowledge and the mind generally in mainstream anglophone philosophy for more than a century. That it nevertheless retains a vestige of dualism seems to me emblematic of a deep problem with that framework, but this larger point cannot be defended here. The broad message offered here is that the Aristotelian idea of habituation is well supported in experimental psychology, which requires philosophers of mind to take the temporal aspect of agency more seriously than it is taken by Moran and the discussions of rationality and self-knowledge with which he engages.

My argument concentrates not on knowledge of one's own beliefs, but on knowledge of one's own desires. Moran's focus is on belief, but he considers his account to cover the cases of desire and intention too. The problem posed to Moran's account by habituation is clearest and best supported by empirical psychology in the case of desires. Having established this problem and traced it back to an implicit mind-body dualism in Moran's account, I will present some evidence that a parallel problem arises for belief states. But this is not essential to my argument. For if we are to understand first-person authority in the context of ourselves as agents rather than simply as knowers, then the problem posed to Moran's account by the reality of desire is enough to require a fresh approach. Before raising these problems, however, I will outline Moran's account of the nature and limits of first-person authority.

1. Moran's Account of First-Person Authority

Moran's central point is that first-person authority is essential to rational agency. As rational agents, he argues, our minds are responsive to our own deliberations. My beliefs are formed by my consideration of the objects of those beliefs. The authority of the declaration 'I believe that p' is not grounded in some special access to the prior fact that I believe that p. It is grounded in the deliberation over whether p is true determining my doxastic state with respect to p. There is a single judgment that can be expressed either as 'p' or reflectively as 'I believe that p' and this judgment ordinarily causes the formation in me of the belief that p. This is why we have this authority only over our mental states, he argues, and not over the mental states of other people or over other aspects of ourselves such as our health (2001: 32-5). This is also why first-person authority is not merely a useful way of gaining information, one that might one day be replaced by some more accurate technology, but is rather an ability whose functioning is essential to the agent's psychic health (2001: 89-94, 148-50).

This is not the only form of self-knowledge, according to Moran. One can also come to know of one's mental states through inference from observation of one's words and deeds. But this form of self-knowledge does not entail first-person authority. Anyone with access to the relevant observational information could draw the same conclusions. One might typically have access to more of this information than other observers have, because usually one is the only constant witness to one's words and deeds. But this is merely a contingent advantage, one that could be cancelled out by memory problems, motivations to see oneself in a positive light, or surveillance technology. In principle, this form of access to one's mental states is no less available to other people than it is to oneself. Moreover, an agent who was entirely reliant on such evidence to find out about their own mental states could not treat their mind as their own to make up. Such an agent could not deliberate or form a judgment, so could not be fully rational. Thus, the transparency of 'do I believe that p?' is not merely a descriptive fact about human beings, but 'a normal rational *expectation* we make of them' (2001: 68).

Moran does allow that local failures of first-person authority occur within the global context of a well-functioning rational agent. In such cases, the form of self-knowledge grounded in observation takes precedence. Moran gives an example from psychoanalysis. Someone 'might become thoroughly convinced, both from the constructions of the analyst, as well as from her own appreciation of the evidence' that the attitude that she has been betrayed by a sibling 'must indeed be attributed to her', even though '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' (2001: 85). Deliberation over whether p, in this case, does not lead the agent to the conclusion that p. The ordinary agential route to answering the question whether she believes that p, therefore, leads her to conclude that she does not. Yet the observational evidence shows that she does believe that p.

If this reasoning is right, it provides a tight and sophisticated explanation of some of the traditional basic claims of psychoanalysis. It explains what it means for a belief to be beyond the reach of ordinary self-knowledge without undermining the generally authoritative status of that form of self-knowledge. It explains how such a belief can nevertheless be uncovered through the therapeutic procedure. And it explains why the agent whose belief it is might not only be surprised to find that they have the belief, but might sincerely deny having it. They might do so because when they consider whether they believe that p, they quite naturally fall into the usual procedure of considering whether p. In the example Moran gives, the woman's denial of having the belief that her sibling betrayed her is the sincere result of considering whether her sibling betrayed her. This is why she needs to be presented with the evidence that she does have the belief, rather than simply have the conclusion drawn from that evidence announced by her analyst.

What implications does this explanation of psychoanalytic practice have for normal rational functioning? If this agent reasons to the conclusion that she has not been betrayed by her sibling, she is able to correctly report 'I believe that I have not been betrayed by my sibling'. But this leaves in place, according to Moran's explanation of the case, another belief whose content contradicts that of the belief she has reported. She correctly declares 'I believe that not-p', yet continues also to believe that p. For this to be a failure of normal functioning, therefore, requires more than that the judgment that not-p is normally the formation of a belief that not-p. It requires also that in normal functioning this belief displaces the agent's previous belief on the matter. Indeed, this would seem to be a requirement of rationality: coming to believe something ought to displace any prior belief that directly contradicts the new belief. What has failed in this psychoanalytic case is not belief-formation itself, but rather the expected rational effect of such belief-formation on the agent's overall doxastic state.

2. Deliberation and the Displacement of Desire

Should we accept that, in normal psychological functioning, the formation of a new belief through deliberation displaces any previous belief that directly contradicts it? One aim of this paper is to show that there is empirical evidence against this picture of belief formation and revision, even though it is presupposed by much recent anglophone philosophy. As an account of normal psychological functioning, this picture is indebted to a dualism of the rational and the arational that current experimental psychology suggests is false. In order to reach this conclusion, we will first consider the case of desire. Moran does claim that his account of first-person authority covers knowledge of one's own desires as well as knowledge of one's own beliefs and many of his examples are cases of desire. Despite this, his theoretical analysis remains closely focused on knowledge of one's own beliefs, with only a few pages devoted explicitly to applying the theory to the case of desire.

This is important because the idea that the deliverance of deliberation is a new mental state that displaces any that directly contradict it faces an obvious difficulty in the case of desire. It is quite common for a desire to persist in opposition to the outcome of deliberation. Where we then act on the desire and against our judgment, this is a case of akrasia or weakness of will. But cases where we go on to act on the judgment and against the recalcitrant desire present the same problem for Moran's account. For the problem is the persistence of the desire itself, not its manifestation in action. Moreover, these kinds of case are not restricted to the kinds of failure of first-person authority that can be uncovered through psychotherapy. We are often very well aware that we want to do something other than what we judge to be the best thing to do. And our awareness of this does not seem to depend on observation of evidence that would be directly available to another person. This is why such cases are often depicted in terms of an inner struggle with oneself, the outcome of which is often described using the metaphors of weakness and strength.

Moran's response to this difficulty is to distinguish desires that are sensitive to deliberative judgment from those that are not. The former, which he calls 'motivated' or 'judgment-sensitive' desires, have not necessarily been produced by reasoning. They are categorised together purely as a result of their being responsive to reasoning about their objects. These are the desires that can be straightforwardly displaced or refined by a deliberative judgment. A motivated desire is justified by a set of beliefs about the desired object in such a way that losing that justification should lead to the loss of the desire. 'It is the normal expectation of a person', writes Moran, 'as well as a rational demand' that 'the question of what he actually does desire should be dependent in this way on his assessment of the desire and the grounds he has for it' (2001: 115). His example is a desire to change jobs, which depends for its justification on beliefs about oneself, one's current job, and prospects for other employment. If these beliefs are revised in light of new evidence, then there rationally ought to be, and we usually expect there to be, a corresponding revision of the desire to change jobs.

Moran contrasts these with a judgment-insensitive kind of desire, which he labels 'brute desire' (2001: 115, 116). His examples are desires 'associated with hunger or sheer fatigue' (2001: 114), 'desires of hunger and lust' (2001: 116), and perhaps also 'mere feelings, including such things as the sensation of thirst' (2001: 116). These are not the result of deliberation and neither are they sensitive to deliberation. 'Like an alien intruder, they must simply be responded to, even if one doesn't understand what they're doing there or what the sense of their demands is' (2001: 114-5). It is not true to say that we have no control over these kinds of desire. Rather, the kind of control that we do have is not a directly rational control. One can produce these kinds of desires 'by training, mental discipline, drugs, the cooperation of friends, or simply by hurling himself into a situation that will force a certain response' (2001: 117). This kind of control, which can also be exercised over judgment-sensitive desires, is not entirely independent of rationality. One might adopt such strategies for good reasons. Nevertheless, the operation of the strategy itself would not be a rational operation.

This is a neat response to the problem posed to his theory of first-person authority by the familiar phenomenon of desires that persist in opposition to the deliverances of deliberation. For not only would it isolate the problem within a particular category of desires, which Moran considers to play only a minor role in overall behaviour (2001: 116). It would also explain why these recalcitrant desires do not exemplify the kind of failure of first-person authority that can be identified only with the help of an impartial observer of the details of one's behaviour, such as a psychoanalyst. Given the kind of mental state a belief is, on Moran's view, the failure of a belief to be displaced by deliberation resulting in a directly contradictory judgment is a failure of normal rational functioning. But it is not constitutive of a 'brute desire' that it should be sensitive to deliberation, so this is beyond the boundaries of normal first-person authority rather than a malfunction. Moran does not address the question of how we know about these desires, given that our experience is not generally one of inferring them from our behaviour. But there seems no obvious reason why this could not be answered consistently with Moran's theory of our knowledge of our own judgment-sensitive mental states.[1]

3. Habituation of Deliberative Desire

What would pose a significant problem, however, would be good reason to reject the identification of normally recalcitrant desires, those that are not displaced by a contradictory deliberative judgment, with 'brute' bodily desires. If normal psychological functioning renders desires that were originally formed through deliberation resilient in the face of contradictory judgment, then such resilience could not be understood as restricted to cases of 'brute' desire and cases of malfunction to be diagnosed through psychoanalysis. Moran's view is that our epistemic authority over our own mental states rests on our ability to set those states by deliberation. Where deliberative judgment does not change a mental state it directly contradicts, either this mental state was always beyond the normal reach of deliberative influence or something has gone wrong. Our own experience, however, provides us with good reason to think that a desire arrived at by deliberation can become habituated through ordinary functioning to such a degree that the kind of simple rational revision that Moran's theory requires is no longer possible. Moreover, as we will see in the next section, this picture of habituation is supported by the leading model of the consistencies and variations in an individual's behavioural cognition.

Habituation of a mental state has been understood at least since Aristotle primarily in terms of its repeated employment in reasoning resulting in action (NE: 1105b9-18). Moran is right to say that if one learns new information about the development of one's current job or one's prospects for other employment, then a recently formed desire to change jobs will be rationally sensitive to this new information unless one's cognition is not functioning as it should. But compare this to the desire to pursue a particular highly

[1] For further consideration of this question, see Webber 2015b: §§ 2 and 5.

competitive career that requires significant qualifications. One may have decided many years ago that one wants to be an academic philosopher, for example, and employed this desire regularly in making many large and small decisions, including ones concerning a range of financial and personal risks and sacrifices over a sustained period of time. If one then learns, towards the end of this time, that one is unlikely to settle into a stable academic job for many more years after doctoral graduation than one had realised, or that academic life involves far more bureaucracy than one had imagined, then these considerations might lead one to decide that a new direction would be better. But this decision is unlikely to simply displace the old desire. If one does take a new career path, this is likely to be haunted for some time by the old desire. Or one might still pursue the original career despite one's reflection that this no longer seems so wise, given one's other life goals.

Might this case be understood in terms of the distinction between the object desired and the aspects of that object that make it desirable? Could we say that the desire for the academic career as a whole is rationally revised by the judgment, even though the research and teaching continue to be found just as desirable? This move does not capture all the ways that this kind of case might develop. Someone who gives up on an academic career for these reasons might nevertheless later experience envy at the career of a friend from graduate student days who has continued into academia. This would not require forgetting about or reevaluating the original reasons for changing career. The envy might be accompanied by the judgment that leaving academia was the right decision. It is simply that this judgment does not nullify the desire for the academic career. Conversely, in the case where one continues with the original career path, it may be the case that the new discoveries about the profession would have deterred one a few years earlier without there being any change over that time in the degree to which one enjoys or values teaching and research or in the way one would evaluate these new discoveries. All that needs to have changed is the degree to which the goal of an academic career, or indeed of the goals of research and teaching, have become embedded in one's outlook. The more ingrained such a goal is, the greater the force a countervailing consideration would require to remove it.

Careers are not the only cases where a desire can feature pervasively in one's practical reasoning for many years. This extreme kind of habituation is perhaps more commonly experienced in personal relationships. But these kinds of cases should not simply be considered as isolated potential counterexamples to Moran's thesis. For the way that they operate indicates an aspect of all desire that presents a deep problem for the idea that in normal functioning a mental state is displaced by a directly contradictory deliberative judgment. These cases exemplify the power of habituation in an extreme and hence easily noticeable way. Since habituation is a matter of degree, however, we should expect this resilience in the face of contradictory judgment to be a matter of degree too. What is more, we seem to be aware of these kinds of desires in a way that does not rely on observation of our own behaviour. This kind of self-knowledge, therefore, is not either of the two kinds that Moran describes. The threat this poses to Moran's deliberative account of first-person authority is that all of the work his theory is designed to do might

already be done by the correct account of knowledge of one's own habituated desires. Before considering the evidence from empirical psychology that this is indeed the case, it is worth seeing what has led Moran to an account that faces this problem.

A form of mind-body dualism lies at the heart of Moran's strategy for responding to the problem that desire poses for his theory that first-person epistemic authority rests on the power of deliberation to set the mental state that its conclusion announces. Moran divides desires into a pair of mutually exclusive and collectively exhaustive categories. Those resistant to rational revision are 'brute' desires identified with bodily needs. The others are categorised together with beliefs as directly and immediately responsive to rational deliberation. This matches the Cartesian idea of the mind as the realm of thoughts held together by the rational relations between their informational contents, distinct from the bodily realm that lies beyond its rational reach. Moran is not committed to the metaphysical claim that these realms inhere in distinct kinds of substance, but he is committed to the dualism of the rational and the arational that led Descartes to that metaphysical claim. Moran even implicitly endorses the Cartesian identification of this rational system with the self. He claims that 'brute' desires 'simply assail us with their force' (2001: 116) and are 'mere happenings to which the person is passively subject' (2001: 116). He likens them to intruders coming in from outside (2001: 114-5) and describes strategies for manipulating them as merely 'external means' (2001: 117). However, we should not identify the self, or the agent, with one half of this purported dichotomy between the rational mind and the arational body. For the existence of mental states that have been formed and habituated through reasoning but which, having become habituated, are no longer immediately responsive to reason shows that this is a false dichotomy.

4. Evaluative Attitudes in the Personality System

Habituation is central to the 'cognitive-affective system theory' of personality. This was developed as a model of the psychological processing that underlies each individual's pattern of behaviour. It is intended to explain why the individual's behaviour in response to a particular feature of their situation will vary with changes in some background features of the situation but not others, or might in some cases be invariant across all such contextual changes. As such, it is supported by a very substantial body of empirical research into this aspect of behaviour (Mischel and Shoda 1995). The essence of the theory is that this cognitive and affective processing should be modelled as a connectionist system. It is not only the set of beliefs and desires that make up the system that matters, but the associative connections between those mental states and the relative strengths of those connections. We should picture the processing itself as a flow of activity through the system, where the activation of one mental state causes the activation of those associated with it to a degree determined by the strength of that connection. Each associative connection is strengthened each time this activity flows along it. This is the role of habituation: repeated use of the same associative connection between two mental states increases the strength of that connection, so increases the proportion of mental activity that flows along it.

This theory is not a complete account of the production of behaviour, but rather a framework for conceptualising further research into the psychology of individual behaviour (Shoda and Mischel 1996: 415). One area of empirical research that lends itself particularly well to this framework is attitude psychology, of which cognitive dissonance theory is the most famous strand. This tradition of social psychology has converged on a conception of an attitude as a cluster of cognitive and affective mental states that together make up an individual's overall evaluation of some object. For example, someone might have an overall attitude of approval towards democracy, made up of the belief that democracy is the best way to keep the peace, the desire that peace be kept, the belief that current Western models of democracy tend to place too much power in the hands of political parties, and so on. Attitude psychology has also converged on the idea that attitudes have strength as well as content. This strength is not the degree of approval or disapproval, which is included in the content. It is rather the degree of influence the attitude has over the agent's cognitive and affective processing.[2] If the constituents of an attitude are seen as elements in the cognitive-affective personality system, then the attitude's strength is given by the number and average strength of the associative connections between them.

Conceptualising attitude psychology this way makes clear why attitude strength correlates with consistency of judgment and action across situations. This is well illustrated by an experiment in which people were asked for their attitude towards Greenpeace and asked some questions designed to measure the strength of that attitude, then a week later given the opportunity to donate to Greenpeace and then asked again about their attitude towards Greenpeace (Holland et al 2002). The attitude content reported at the start of the experiment predicted whether the individual later donated to Greenpeace only where the attitude was strongly held. In these cases, the second report of the attitude generally matched the first. Moreover, where the original attitude was weakly held, the second attitude report was in line with the response to the opportunity to donate even though this did not correlate with the attitude content reported earlier. The experimenters conclude that weak attitudes are subject to situational variation because they are constructed at the time out of whatever relevant beliefs and desires come to mind most easily. If one has just had the opportunity to donate to Greenpeace, then one's knowledge of whether one donated or did not donate is highly accessible and therefore strongly influences one's attitude. But stronger attitudes, comprising a set of strongly interconnected elements in the connectionist personality system, are robust mental states that exert the same significant influence on judgment and behaviour across varying situations.

Where an attitude has been significantly habituated, therefore, it exerts a general pressure towards some outcome in judgment and behaviour. It is, in short, something that philosophers would classify broadly as

[2] For more detailed explanations of attitude psychology, in relation to the philosophical idea of ethical virtue, see Webber 2013 and 2015a.

a desire. But because it is constituted by a complex set of strong associative connections between its constituent mental states, it can be revised or replaced only through progressively weakening those associative connections or strengthening other ones. It will not simply be displaced by the contradictory deliverance of an episode of deliberation. Moreover, the same theory explains why the same strong attitude has less influence over deliberation than over immediate judgment and automatic behavioural cognition. This is essentially because deliberation is a slower process. The stronger an attitude is, the more accessible it is, where this is measured in terms of the speed with which it is brought to bear on any given episode of processing. When the processing itself is completed in a short timeframe, only the most accessible attitudes and other mental states will have any influence. But when the processing takes much longer, many more considerations can be drawn upon. This is how the deliberation that fails to displace a strong attitude can have reached a conclusion contrary to that attitude in the first place.

Thus, attitude psychology and the personality system provide a clear explanation of how desires become habituated and thereby progressively less susceptible to immediate revision or displacement in response to deliberative judgment. Moreover, the weakest attitudes are not susceptible to such displacement either, for they are not persisting states at all. Instead, we should say that the more an attitude becomes a persisting state, the more it becomes a stable part of the individual's cognitive system, the more resilient it becomes to immediate change through deliberation. This is not to say that strong attitudes cannot be revised or displaced, only that doing so is itself a matter of habituation. Therefore, one cannot simply determine one's desire concerning an object by deliberating about that object in the way that Moran describes. The outcome of that deliberation will not immediately become a persisting attitude. It may be at odds with an existing attitude, and if so will not immediately displace it. This is not due to any psychological malfunction. It is rather a feature of our cognitive and affective system that it precludes immediate formation of stable attitudes and conversely produces resilience of stable attitudes to immediate change.

5. One Kind of Implicit Bias as Habituated Belief

Is this structure of habituation paralleled in the case of belief? To show that it is, we would need evidence of beliefs that are not simply displaced when contradicted by judgment. This presents a procedural difficulty. For it is part of the usual philosophical understanding of a belief that it should be immediately responsive to rational deliberation. A mental state that is not judgment-sensitive in this way is likely to be classified by philosophers as something other than a belief. Yet it does seem legitimate to raise the question of whether all belief is like this. My strategy is to provide evidence drawn from the literature on implicit bias that a mental state can behave like a belief in all respects except its sensitivity to deliberative judgment. This same evidence suggests that such a mental state will have been arrived at through the erosion of its judgment-sensitivity through its repeated activation. If this evidence is correct, therefore, it shows that habituation can render a belief beyond direct deliberative control.

The term 'implicit bias' currently covers a disparate range of phenomena, so we should not assume that any significant general account of implicit bias can be provided (Holroyd and Sweetman forthcoming). One phenomenon often placed in this category is the tendency to associate young black men with handguns. This tendency might have been a factor in the death of Mark Duggan, who was killed by an armed police officer in north London in 2011. During the inquest, the officer who shot him claimed that Duggan had been holding a gun at the time and described this scene in detail. Another witness, however, claimed that Duggan had been holding a phone. No gun was found in his possession after he was shot. A gun was found some distance away, but was wrapped in a sock. The inquest jury found that Duggan had been unarmed when he was shot. So how should the police officer's testimony be explained?

A wealth of evidence supports the idea that high-speed decisions about whether someone is armed are biased by whether the person is black. After briefly seeing a black face rather than a white face, people more quickly identify guns as guns and when working at high speed more frequently misidentify other objects as guns (Payne 2001). When playing a video game in which one has to shoot all and only the men carrying guns, people making decisions at high speed shoot black men more frequently than white men, whether armed or unarmed. When playing more slowly, people tend to shoot all and only the armed men irrespective of race, but make their decision to shoot an armed man more quickly when he is black and make their decision not to shoot an unarmed man more quickly when he is white (Correll et al 2002). All of this suggests that people strongly associate black men with handguns in a way that influences the outcome of object identification and decision-making processes executed at high speed, but does not influence their outcomes at low speed (Payne 2006).

Might the armed officer who shot Duggan have misidentified a phone as a gun due to this kind of association? He did have to identify the object and make his decision at high speed. Duggan had been stopped because police had information that he was carrying a gun and this bias can be exacerbated by recent exposure to information linking black men with handguns (Correll et al 2007a). On the other hand, police officers have been found to exhibit this bias only in the time it takes to make the decision whether or not to shoot, differing from the overall population in generally not exhibiting a bias in the content of the decision itself. That this can be replicated in undergraduate students by training them in deciding whether or not someone is holding a gun suggests that police officer training makes a difference here (Correll et al 2007b).

Irrespective of whether this bias was in fact involved in the Duggan shooting, there remains the issue of whether we should think of the cognitive association between black men and handguns as a belief that black men often carry handguns. One recent influential philosophical discussion of this kind of implicit bias argues that we should not. Instead, according to Tamar Gendler, we should think of this association as an 'alief', which is an arational state that can be had by other animals (2008: 557, 574). Why should we agree that the mental state is arational? It does seem to rationalise its cognitive and behavioural effects. If

it were true generally that a black man is likely to be carrying a handgun, then that would rationally support the expectation of a handgun that influences perception and decision in each case. Gendler denies this rational relation on the grounds that the agent would deny having the requisite belief (2008: 565). But this shows only that if the state is a belief that rationalises its effects, then it might be one that the agent is unaware of having.

Gendler's central reason for denying that the mental state is a belief is that it is not sensitive to evidence in the right way. Belief, she claims, 'is normatively governed by the following constraint: belief aims to "track truth" in the sense that belief is subject to immediate revision in the face of changes in our all-things-considered evidence' (2008: 585). Why should we accept this denial of the possibility of habituated belief? Gendler's reason is that belief has to be rationally sensitive in this way 'if it is to bear the relation to knowledge and rationality that philosophers require of it' (2008: 563). But this argument can be reversed: if the empirical evidence shows that belief can be habituated in a way that prevents immediate displacement by directly contradictory deliberative judgment, then any philosophical theory that rests on the denial of such habituated belief will need to be rejected.

Not only does the mental state underlying this form of implicit bias rationalise its psychological and behavioural effects, but it is also formed in a way that rationally tracks the individual's experience. For example, police officers working in communities with both a high crime rate and a high proportion of black residents showed a particularly strong bias in the time it takes to decide whether or not to shoot a character in the video game (Correll et al 2007b: 1021). In the general population, the bias can be magnified temporarily by increasing the proportion of black characters in the video game who are carrying guns (Correll et al 2007a: Study 2). Moreover, the bias is strongly correlated with the individual's knowledge that the cultural stereotype of black men associates them with violence, irrespective of whether the individual endorses that stereotype (Correll et al 2002). Since this cultural stereotype is propagated through media imagery across news stories, fictional stories, and music lyrics and videos, knowledge of this stereotype is likely to be due to exposure to this imagery.

The mental state underlying this bias, therefore, shares with paradigmatic cases of belief both its being rationally supported by the individual's own experience, however unrepresentative of reality that experience may be, and it in turn rationalising its psychological and behavioural effects. It is different only in not being immediately sensitive to deliberative judgment that contradicts it. This suggests that receptivity to information that contradicts one's overall experience is limited, since beliefs formed on the basis of that experience are likely to be deeply ingrained. Although it would be more rational to discount the experience of media that one knows to significantly distort reality, it seems that normal psychological functioning precludes the influence of this distortion being counteracted easily. This would explain why people who repudiate the cultural stereotype associating black men with handguns still manifest the bias rationalised by that stereotype. Thus, rather than classify the mental state underlying this bias as an alief,

we should consider it a belief that is not revised immediately by any judgment that contradicts it. Since it has been formed through repeated exposure to the stereotype, this recalcitrance seems due to habituation.

6. Knowing One's Own Habituated Beliefs

This analysis of one form of implicit bias as rooted in habituated belief leaves open the question of whether this habituation operates through practical reasoning, as the Aristotelian position suggests, or whether repeated exposure to the cultural stereotype is sufficient. If it does require reasoning, this need not be explicit deliberation, for the background processing of information required for following a news story or fictional narrative is also a form of practical reasoning. But even if this kind of habituated belief comes about through mere exposure to the stereotype, this would at most show that the Aristotelian picture is incomplete, and that reasoning is not necessary for habituation. The same question arises for the case of desire. But consideration of the experimental research into the effect of mere exposure on attitudes and other cognitive states must await another occasion.

Our primary concern here is with the implications of habituated belief, however it comes about, for self-knowledge and first-person authority. The common cognitive association of black men with handguns is routinely described in the philosophical and psychological literature as beyond the scope of ordinary self-knowledge, often as 'unconscious' or 'not available to introspection'. The term 'implicit' has come to be used in this sense, even though it originally labelled only a style of measuring mental states. (The phrase 'implicit association test' denotes an implicit test of associations, not a test of implicit associations.) But this classification generally occurs without any serious consideration of what it means. It is usually motivated by the claim that the subjects explicitly stated that they do not think that black men are strongly associated with handguns. But if these statements represent deliberative judgments about the relation between black men and handguns, then we should not be so quick to assume that they express ordinary self-knowledge of the speaker's belief states. For the idea that ordinary knowledge of one's own beliefs is grounded in deliberative judgment about the objects of those beliefs assumes that beliefs do not become habituated as they become stable mental states, which we can now see is not a safe assumption to make.

Might there be another way in which deliberation provides knowledge of one's own habituated beliefs? To deliberate requires one to draw on relevant information. There is evidence that a belief that has been habituated through repetition of some claim will be more easily and rapidly accessible to deliberation than one that has not. Even when the subject has explicit reason not to believe that the claim is correct, this habituation leads to the automatic retrieval and application of the claim, which can be counteracted only through deliberation drawing on the reason not to believe the information (Begg et al 1992). A belief habituated through exposure to distorted media could therefore be brought automatically to deliberative cognition even though it might then be deliberatively defeated by the knowledge that the media is distorted in this respect. If this is right, then one could know one's own habituated beliefs through their

regularly and rapidly coming to mind in deliberation. But this would not entail that all our habituated beliefs can be known in this way. For it would leave open that a habituated belief might be counteracted by another habituated belief in the automatic cognition that subserves deliberation, such that its content never attains the status of being a consideration in deliberation.

This is a speculative suggestion, full consideration of which would require further investigation of the vexed issue of the role of habituated belief in the relation between automatic and deliberative cognition (Thompson 2009). But it does seem possible that belief might mirror desire in that the more habituated it is, the more stable it is as a persisting feature of the subject's cognitive system and the more resistant it is to being changed by a contradictory judgment. If this is right, then it undermines the idea that first-person epistemic authority over one's beliefs consists in one's agential authority to form those beliefs. For not only would it show that we lack such authority over strongly habituated beliefs, but it would also show that deliberative judgment does not in itself produce persisting beliefs at all. In the absence of habituation, such a judgment just reflects the considerations that come to mind in that deliberative process. The considerations that come to mind soonest and most easily would in general be those that are themselves most strongly habituated, but on any given occasion these would be joined by any considerations that have very recently been thought about. Moreover, the full range of considerations that come to mind will depend on the length of time devoted to the deliberation. One is likely, therefore, to reach divergent judgments on the same topic on different occasions. If this is right, then to report a deliberative conclusion is not to report a persisting belief.

7. The Temporal Dimension of Rational Agency

Moran's theory of first-person authority as the deliberative agential authority to make up one's own mind therefore rests on a dualism of the rational and the arational that does not allow for habituation. We have good intuitive and empirical reason to accept that habituation is central to the normal psychological functioning of desire. There is some empirical support for the idea that habituation plays a parallel role in belief. If there is to be genuine first-person epistemic authority over persisting mental states, therefore, an alternative account to Moran's is required in the case of desire and might also be required in the case of belief. Ought such an account respect the idea of transparency? The same considerations that we have raised against Moran's account of first-person authority also show that his deliberative explanation of transparency applies only where the belief or desire reported is not a persisting mental state. For to reach a conclusion by deliberation is not in itself to form a persisting mental state, and neither does the conclusion displace any contradictory mental state as a matter of normal psychological functioning. If transparency can be understood in some other way, it could be a feature of genuine first-person authority over persisting mental states. But since the intuitive appeal of the idea of transparency can be explained by its relation to transient beliefs and desires, we should not assume that first-person authority over persisting mental states has anything to do with transparency.

The rejection of dualism entailed by the recognition of habituation need not involve a wholesale rejection of Gendler's conception of alief. It requires only that a mental state not immediately sensitive to rational judgment is not thereby entirely outside the realm of reason. We have seen that the mental state underlying one form of implicit bias seems to be rational in both its formation and its effects while being no longer immediately revisable through judgment. Moreover, Gendler describes one kind of alief as being rationally sensitive to statistical evidence in sophisticated ways that are not often matched by deliberation (2011: 33-36, 54-57). The problem here is the dualistic opposition of rational and arational, which Gendler maps onto the difference between humans and other animals. We should instead accept that the rationality of a mental state is a matter of degree and, as a result, the rationality of a creature is too. An animal whose mental states are sensitive to the environment in statistically sophisticated ways is more rational than one whose mental states are not, irrespective of whether that first creature is also capable of deliberative judgment.

Recognition of the role of habituation in our psychology clarifies the sense in which we are rational animals, or imperfectly rational agents. It is not that we are rational angels unfortunately yoked to mortal bodies that interfere with our rational processes by assailing us with their needs and demands. Neither is it that our rational systems are simply overlain on the associative alief mentation of our animal bodies that exerts its own motivational pressures. It is rather that our form of rationality itself inherently involves a kind of habituation that limits our deliberative control over desires, and perhaps also beliefs. This habituation operates through the repeated employment of a mental state in reasoning, whether as a conclusion or as a premise Moreover, it is rational for a cognitive system of finite capacity to rely on habituation in this way. Progressively embedding a deliberative premise or conclusion in the cognitive system as it is repeatedly employed obviates the need to continue to revisit each of these chains of reasoning in order to consider one's commitment to it, but does so in a way that is somewhat sensitive to its degree of rational support without needing to record that support itself in an accessible format. This is an efficient design.

To put this another way, due recognition that the dualism of the wholly rational and the wholly arational is false requires recognition that our form of rationality itself is essentially temporal. Our form of rationality relies on habituation, which is a process that requires significant stretches of time. Our form of rationality draws on such habituated mental states, which have been formed through past rational processes to a degree that determines their influence over cognition and their resistance to immediate change even though the contents of those rational processes may be long forgotten. This temporal aspect of rational agency needs to be borne in mind when considering how one might aim to remove some habituated belief or desire, or at least prevent its manifestation in action. Manipulating one's environment in order to counteract the distortions of media representation, such as by putting pictures of counter-stereotypical individuals in prominent places for example, should not be thought of as a merely causal and arational way

of changing one's mind, since it operates through the chronic rational sensitivity of habituation. Conversely, the possibility of changing one's habituated mind through deliberative means, such as discussing the reasons for one's decisions or making decisions sufficiently slowly to regularly employ beliefs and desires that are less accessible, should not be dismissed lightly.

Philosophical and empirical consideration of the extent of our deliberative control over our own minds should be explicitly framed by this point about the essentially temporal nature of our rationality, as indeed should consideration of all forms of agential control over the contents of minds. Likewise, further discussion of our epistemic access to our own minds and those of other people, in relation to our deliberative capacities and more generally, should keep this temporal dimension of human rationality sharply in focus. We are not simply abstract reasoners. We are essentially temporal rational agents. We should keep reminding ourselves of that until the time comes when its significance is automatically taken into account.[3]

[3] This paper was developed through presentations at Cardiff University work-in-progress seminar, Kings College London philosophy society, the visiting speaker seminar at Manchester Metropolitan University, and the 2013 conference of the Nordic Society for Phenomenology at University of Copenhagen, and through participation in the Implicit Bias Project at University of Sheffield. I am grateful to the organisers and participants of these for helping to shape my thoughts on this issue. I am also very grateful to Roman Altshuler, Jules Holroyd, and Michael Sigrist for their thoughtful responses to the first draft.

Bibliography

Aristotle. NE. *Nicomachean Ethics*. Translated by Christopher Rowe. Introduction by Sarah Broadie. Oxford: Oxford University Press, 2002.

Begg, Ian Maynard, Ann Anas, and Suzanne Farinacci. 1992. Dissociation of Processes in Belief: Source Recollection, Statement Familiarity, and the Illusion of Truth. *Journal of Experimental Psychology: General* 121: 446-458.

Correll, Joshua, Bernadette Park, Charles Judd, and Bernd Wittenbrink. 2002. The Police Officer's Dilemma: Using Ethnicity to Disambiguate Potentially Threatening Individuals. *Journal of Personality and Social Psychology* 83: 1314-1329.

Correll, Joshua, Bernadette Park, Charles Judd, and Bernd Wittenbrink. 2007a. The Influence of Stereotypes on Decisions to Shoot. *European Journal of Social Psychology* 37: 1102-1117.

Correll, Joshua, Bernadette Park, Charles Judd, Bernd Wittenbrink, and Melody Sadler. 2007b. Across the Thin Blue Line: Police Officers and Racial Bias in the Decision to Shoot. *Journal of Personality and Social Psychology* 92: 1006-1023.

Gendler, Tamar Szabó. 2008. Alief in Action (and Reaction). *Mind and Language* 23: 552-585.

Gendler, Tamar Szabó. 2011. On The Epistemic Costs of Implicit Bias. *Philosophical Studies* 156: 33-63.

Holland, Rob W., Bas Verplanken, and Ad van Knippenberg. 2002. On the Nature of Attitude-Behavior Relations: The Strong Guide, The Weak Follow. *European Journal of Social Psychology* 32: 869-876.

Holroyd, Jules, and Joseph Sweetman. forthcoming. The Heterogeneity of Implicit Bias. In *Implicit Bias and Philosophy*, edited by Michael Brownstein and Jennifer Saul. Oxford: Oxford University Press.

Mischel, Walter and Yuichi Shoda. 1995. A Cognitive-Affective System Theory of Personality: Reconceptualizing Situations, Dispositions, Dynamics, and Invariance in Personality Structure. *Psychological Review* 102: 246-268.

Moran, Richard. 2001. *Authority and Estrangement: An Essay on Self-Knowledge*. Princeton: Princeton University Press.

Payne, B. Keith. 2001. Prejudice and Perception: The Role of Automatic and Controlled Processes in Misperceiving a Weapon. *Journal of Personality Social Psychology* 81: 181-192.

Payne, B. Keith. 2006. Weapon Bias: Split-Second Decisions and Unintended Stereotyping. *Current Directions in Psychological Science* 15: 287-291.

Shoda, Yuichi and Walter Mischel. 1996. Toward a Unified, Intra Individual Dynamic Conception of Personality. *Journal of Research in Personality* 30: 414-428.

Thompson, Valerie A. 2009. Dual-Process Theories: A Metacognitive Perspective. *In In Two Minds: Dual Process Theories and Beyond*, edited by Jonathan Evans and Keith Frankish. Oxford: Oxford University Press.

Webber, Jonathan. 2013. Character, Attitude and Disposition. *European Journal of Philosophy*. DOI: 10.1111/ejop.12028

Webber, Jonathan. 2015a. Instilling Virtue. In *From Personality to Virtue: Essays in the Ethics and Psychology of Character*, edited by Alberto Masala and Jonathan Webber. Oxford: Oxford University Press.

Webber, Jonathan. 2015b. Knowing One's Own Desires. In *Philosophy of Mind and Phenomenology: Conceptual and Empirical Approaches*, edited by Daniel Dahlstrom, Andreas Elpidorou, and Walter Hopp. New York: Routledge.