Implicit bias, awareness and imperfect cognitions

Jules Holroyd

Department of Philosophy, The University of Nottingham, University Park, Nottingham NG7 2RD, United Kingdom

Abstract

Are individuals responsible for behaviour that is implicitly biased? Implicitly biased actions are those which manifest the distorting influence of implicit associations. That they express these ‘implicit’ features of our cognitive and motivational make up has been appealed to in support of the claim that, because individuals lack the relevant awareness of their morally problematic discriminatory behaviour, they are not responsible for behaving in ways that manifest implicit bias. However, the claim that such influences are implicit is, in fact, not straightforwardly related to the claim that individuals lack awareness of the morally problematic dimensions of their behaviour. Nor is it clear that lack of awareness does absolve from responsibility. This may depend on whether individuals culpably fail to know something that they should know. I propose that an answer to this question, in turn, depends on whether other imperfect cognitions are implicated in any lack of the relevant kind of awareness.

In this paper I clarify our understanding of ‘implicitly biased actions’ and then argue that there are three different dimensions of awareness that might be at issue in the claim that individuals lack awareness of implicit bias. Having identified the relevant sense of awareness I argue that only one of these senses is defensibly incorporated into a condition for responsibility, rejecting recent arguments from Washington & Kelly for an ‘externalist’ epistemic condition. Having identified what individuals should – and can – know about their implicitly biased actions, I turn to the question of whether failures to know this are culpable. This brings us to consider the role of implicit biases in relation to other imperfect cognitions. I conclude that responsibility for implicitly biased actions may depend on answers to further questions about their relationship to other imperfect cognitions.

© 2014 Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/3.0/).

1. Introduction

Are individuals responsible for behaviour that is implicitly biased? Implicitly biased actions are those which manifest the distorting influence of implicit associations. That they express these ‘implicit’ features of our cognitive and motivational make up has been appealed to in support of the claim that, because individuals lack the relevant awareness of their morally problematic discriminatory behaviour, they are not responsible for behaving in ways that manifest implicit bias. However, the claim that such influences are implicit is, in fact, not straightforwardly related to the claim that individuals lack awareness of the morally problematic dimensions of their behaviour. Nor is it clear that lack of awareness does absolve from responsibility. This may depend on whether individuals culpably fail to know something that they should know. I propose that an answer to this question, in turn, depends on whether other imperfect cognitions are implicated in any lack of the relevant kind of awareness.

This article is part of a special issue of this journal on Imperfect Cognitions.

E-mail address: Jules.Holroyd@Nottingham.ac.uk

http://dx.doi.org/10.1016/j.concog.2014.08.024

1053-8100/© 2014 Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/3.0/).
In Section 2 clarify our understanding of ‘implicitly biased actions’ and then argue that there are three different dimensions of awareness that might be at issue in the claim that individuals lack awareness of implicit bias. Having identified the relevant sense of awareness, in Section 3, I argue that only one of these senses is defensibly incorporated into a condition for responsibility, rejecting recent arguments from Washington & Kelly for an ‘externalist’ epistemic condition. Having identified what individuals should – and can – know about their implicitly biased actions, I turn in Section 4 to the question of whether failures to know this are culpable. This brings us to consider the role of implicit biases in relation to other imperfect cognitions. I conclude that responsibility for implicitly biased actions may depend on answers to further questions about their relationship to other imperfect cognitions.

2. Implicit bias, and awareness of it

What are the phenomena at issue when we talk of implicit biases? Amodio and Mendoza (2010) describe them as ‘associations stored in memory’ (364). These associations can influence behaviours and judgements. For example, implicit associations have been posited as explaining differential evaluations of the same CV whose only difference was race, indicated by the name at the top (Dovidio & Gaertner, 2000); as implicated in shooter bias, whereby in a computer simulation individuals were more likely to ‘shoot’ black men with weapons than white men with weapons (Glaser & Knowles, 2008); and as playing a role in the seating distances between experimental participants and stigmatised group members (Tidswell, Sheeran, & Webb, in preparation). But what is it about the associations involved in producing such varied behaviour that makes them implicit, and how do we delineate which of the many associations of this sort constitute biases?

2.1. ‘Implicit’

Some have suggested that associations are implicit simply because the measure used to access them is an implicit one; namely, one that does not rely on self-report measures, nor the voluntary offering of information about one’s attitudes. An implicit measure might involve a prime of which the agent is not aware, then a measure of how being so primed influences behaviour or judgement. But why use an implicit measure to access these associations? One reason is that individuals may not be forthcoming or frank about associations they would rather they did not have. Another reason is that the associations are characterised by features of automatic processes which render them difficult for the agent to identify and report on. DeHouwer, Teige-Mocigemba, Spruyt, and Moors (2009) pick out the following features as ones taken to be characteristic of implicit associations: operation without the guidance of proximal goals (that would enable the agent to initiate, intervene or stop the processes); operation without substantial cognitive resources (such as when one’s attention is occupied with some other task); and operation with very limited time (such as when one is required to respond very quickly); or without awareness. Notably, some philosophers and psychologists have taken this latter feature as characteristic or even definitional of implicit bias (see e.g. Kelly & Roedder, 2008; Washington & Kelly, in press; Greenwald & Banaji, 1995; Saul, 2013). I will say much more about this characteristic in the following. What is important is that these features are not specified as necessary for an association to be implicit, which leaves considerable scope for variation in the properties of the associations being measured in such studies, and discussed in subsequent philosophical literatures.2

2.2. ‘Bias’

What is it about some implicit associations that should lead us to characterise them as ‘biases’? We can think of such associations as biases when they are disposed to exert a distorting influence on judgement. The influence at issue can be characterised as distorting in that it leads to a judgement which departs from the norms of rationality. This can be most clearly seen in the CV studies mentioned above: the name at the top of a CV does not provide a reason for judging it to be better or worse than an otherwise identical CV. It is less clear how this analysis explains the behavioural outputs, such as increased seating distance from stigmatised groups. One possibility would be to extend the definition to include not only distortions of judgement but also undesired or undesirable influences on action. Another would be to suppose that these behaviours are preceded by (tacit) judgements, which are distorted (judgements about the suitable place to arrange the seat say; or about the level of danger posed by an individual). Perhaps either way of proceeding is adequate, but for the sake of providing a simple contrast with explicit bias, I will work with the model that sees all implicitly biased action as involving a distortion of judgement. This judgement is sometimes the output measured; in other cases it informs the behavioural output which is measured.3

We can proceed, then, with the following understanding of implicit bias: it is operative when implicit associations produce a distorting influence on judgement and hence behaviour informed by that judgement (this leaves room for some implicit associations which are not implicit biases).

1 As to what associations are – what ontology of the mind best accommodates them – I remain agnostic. Nothing in my argument to follow depends on one particular interpretation of what implicit associations are or how they are structured.
2 Thanks to Robin Scaife and Tom Stafford for very helpful discussion of how we should understand the notion of ‘implicit’ in this context. This passage has been informed not only by these discussions, but by Robin’s very useful blogpost on this topic: http://blogs.nottingham.ac.uk/biasandblame/2014/02/24/what-is-implicit-about-implicit-biases/#comments.
3 Again, many thanks to Tom Stafford and Robin Scaife for fruitful discussions of this issue.
2.3. Responsibility

What are we asking when considering whether or not an agent is responsible for implicit bias? One dimension of responsibility is forward-looking: is this something an agent can be asked to take responsibility for, and bring about changes in her cognition? This is an important sense of responsibility when considering implicit biases, where one of the primary aims is to bring about that individuals act in ways that are less biased. But this is not the only sense of responsibility I am concerned with here. The question here is whether an individual can be held responsible, in the sense of liable to praise or blame, for the manifestation of implicit bias. If an individual acts in ways that express bias, is this something that they are blameworthy for (there is, of course, a further question about whether expressing blame would be appropriate)? To say that the agent is blameworthy, then, is to say that they have intentionally done something that violated a moral standard that we expected them to maintain, and as a result certain responses would be warranted: disapprobation or other forms of informal sanction on the part of others; resentment on the part of the wronged party; guilt on the part of the wrong-doer, and resolution to avoid such behaviours or actions in future (indeed, to take responsibility for that). Whether individuals are responsible in this sense is the question with which I am concerned.

2.4. Awareness

Authors who have addressed the question of responsibility for implicit bias have argued that to the extent that individuals are not (Saul, 2013) or could not reasonably be expected to be (Washington & Kelly, in press) aware of their implicit biases, they are not responsible for action influenced by these biases. However, there are different senses of ‘awareness’ at work in this debate. In the next sub-section I articulate three different senses of awareness that are circulating in the literatures (philosophy and psychology) about implicit biases. This will enable us to identify more precisely the sense in which individuals have or lack awareness, such that we can evaluate whether being in such an epistemic situation exculpates from moral responsibility.

2.5. Three kinds of awareness

What do individuals lack awareness of, in the case of implicit bias? In the literature from empirical psychology and from philosophy, we find different views on this. For example, in making the claim that individuals should not be blamed for implicit biases, Saul writes that ‘a person should not be blamed for an implicit bias of which they are completely unaware’ (Saul, 2013 p. 55), where what is at issue is that individuals are not aware of the operation of the bias in the production of action. We might also interpret Saul as claiming that individuals are not aware of the presence of the bias, but let us restrict our focus to the operation of the bias, in relation to which our concern with responsibility arises most pressingly. The sense of awareness at issue here seems to be introspective awareness; awareness that might yield knowledge of one’s cognitive processes simply by reflecting on one’s internal states and processes. This sense of awareness is also in play in Kelly & Roedder’s description of implicit measures as accessing aspects of cognition ‘not easily accessible or readily available to introspection’ (Kelly & Roedder, 2008, p. 524), and in Anderson’s discussion of ‘unconscious stereotypes’; representations of which the agent is introspectively unaware (2010, p.74, 48). One might have introspective awareness with respect to whether certain beliefs or feelings are playing a role in one’s decisions: one can ask oneself, and on reflection give an answer. But, the claim goes, one cannot simply introspect and discern if an implicit bias is operating in the production of action.

Alternatively, we might be concerned with whether individuals are aware of a set of propositions about implicit bias, which are likely to be true of themselves. This is a second sense of awareness at issue in Saul’s claims: when she writes that individuals may ‘become aware that they are likely to have implicit biases’ (2013, p55), the awareness at issue is of the body of knowledge concerning the disposition of individuals to be biased. Similarly, Washington and Kelly (in press) focus on whether some individuals in fact know, or should know, certain empirical facts about their probable susceptibility to implicit biases. In attempting to explain their divergent intuitions about the responsibility of an egalitarian on a hiring committee who manifests implicit bias in the 1980s, and a similarly placed contemporaneous egalitarian, they observe that ‘in 1980, no one knew the creepy psychological facts about implicit biases; the psychological research had not yet been done, and so today’s wealth of empirical evidence simply did not exist’ (in press). This fact (about what individuals can reasonably be expected to know) figures in their explanation of why they seek to exculpate the 1980s discriminator, but not the contemporaneous one. It is unreasonable to expect the 1980s discriminator to be aware of facts about implicit bias as yielded by empirical psychology: those facts were not part of our epistemic milieu then. But now, those on hiring committees have epistemic responsibilities, which include familiarising themselves with that body of knowledge.

---

6 For an interesting exploration of how individuals might take responsibility, and the social and institutional contexts that might aid doing so in medical care, see Fitzgerald, 2013.

5 This is a plausible and common-sense view of blameworthiness. It is one consonant with Wallace’s articulation and defence of our practices of responsibility (Wallace, 1994).

4 But this is not the only sense of responsibility I am concerned with. The question here is whether an individual can be held responsible, in the sense of liable to praise or blame, for the manifestation of implicit bias. If an individual acts in ways that express bias, is this something that they are blameworthy for (there is, of course, a further question about whether expressing blame would be appropriate)? To say that the agent is blameworthy, then, is to say that they have intentionally done something that violated a moral standard that we expected them to maintain, and as a result certain responses would be warranted: disapprobation or other forms of informal sanction on the part of others; resentment on the part of the wronged party; guilt on the part of the wrong-doer, and resolution to avoid such behaviours or actions in future (indeed, to take responsibility for that).

2.5. Three kinds of awareness

What do individuals lack awareness of, in the case of implicit bias? In the literature from empirical psychology and from philosophy, we find different views on this. For example, in making the claim that individuals should not be blamed for implicit biases, Saul writes that ‘a person should not be blamed for an implicit bias of which they are completely unaware’ (Saul, 2013 p. 55), where what is at issue is that individuals are not aware of the operation of the bias in the production of action. We might also interpret Saul as claiming that individuals are not aware of the presence of the bias, but let us restrict our focus to the operation of the bias, in relation to which our concern with responsibility arises most pressingly. The sense of awareness at issue here seems to be introspective awareness; awareness that might yield knowledge of one’s cognitive processes simply by reflecting on one’s internal states and processes. This sense of awareness is also in play in Kelly & Roedder’s description of implicit measures as accessing aspects of cognition ‘not easily accessible or readily available to introspection’ (Kelly & Roedder, 2008, p. 524), and in Anderson’s discussion of ‘unconscious stereotypes’; representations of which the agent is introspectively unaware (2010, p.74, 48). One might have introspective awareness with respect to whether certain beliefs or feelings are playing a role in one’s decisions: one can ask oneself, and on reflection give an answer. But, the claim goes, one cannot simply introspect and discern if an implicit bias is operating in the production of action.

Alternatively, we might be concerned with whether individuals are aware of a set of propositions about implicit bias, which are likely to be true of themselves. This is a second sense of awareness at issue in Saul’s claims: when she writes that individuals may ‘become aware that they are likely to have implicit biases’ (2013, p55), the awareness at issue is of the body of knowledge concerning the disposition of individuals to be biased. Similarly, Washington and Kelly (in press) focus on whether some individuals in fact know, or should know, certain empirical facts about their probable susceptibility to implicit biases. In attempting to explain their divergent intuitions about the responsibility of an egalitarian on a hiring committee who manifests implicit bias in the 1980s, and a similarly placed contemporaneous egalitarian, they observe that ‘in 1980, no one knew the creepy psychological facts about implicit biases; the psychological research had not yet been done, and so today’s wealth of empirical evidence simply did not exist’ (in press). This fact (about what individuals can reasonably be expected to know) figures in their explanation of why they seek to exculpate the 1980s discriminator, but not the contemporaneous one. It is unreasonable to expect the 1980s discriminator to be aware of facts about implicit bias as yielded by empirical psychology: those facts were not part of our epistemic milieu then. But now, those on hiring committees have epistemic responsibilities, which include familiarising themselves with that body of knowledge.

---

4 For an interesting exploration of how individuals might take responsibility, and the social and institutional contexts that might aid doing so in medical care, see Fitzgerald, 2013.

5 This is a plausible and common-sense view of blameworthiness. It is one consonant with Wallace’s articulation and defence of our practices of responsibility (Wallace, 1994).

6 One might have introspective awareness with respect to whether certain beliefs or feelings are playing a role in one’s decisions: one can ask oneself, and on reflection give an answer. But, the claim goes, one cannot simply introspect and discern if an implicit bias is operating in the production of action.
At issue in these claims, then, is not some particular aspect of one's cognition, nor some observed behavioural effects, but rather some body of knowledge pertaining to individuals' general tendencies to manifest implicit biases. Knowledge of this would be achieved through what we might call inferential awareness; awareness reached through inferences made about this body of empirical knowledge, and one's own behavioural dispositions in light of that.

Finally, other authors are concerned with whether individuals have awareness of the manifestation in behaviour of implicit bias: in particular, their focus is on an individual's awareness of discrepant behavioural responses on tests for biases, and their willingness to attribute such responses to implicit prejudice. For example, Monteith et al. undertook studies on implicit race associations, in which they measured different response times to pairing tasks (black names with unpleasant [congruent] or pleasant [incongruent] terms, and white names with unpleasant [incongruent] or pleasant [congruent] terms. The congruent pairings are those that individuals are expected to respond more quickly on, insofar as they are informed by stronger, more accessible, associations. So, a faster 'black/unpleasant' response than 'black/pleasant' response indicates a stronger association between the former, negative, construct than the latter. Following the study, Monteith et al. 'identified participants who recognized that they were slower on incongruent [contra-implicit association] than on congruent [consistent with implicit association] IAT trials' (2001, p.405), and found that a significant portion of these individuals were able to attribute their differential response times to implicit prejudice. The striking claim here is that a considerable number of individuals (64% of the study participants) were able to recognise, on the basis of observations of their own behavioural responses, that they were responding differently to the different stimuli.

At issue here, then, is whether individuals are aware of their differential or discriminatory behavioural outputs. As a matter of fact, it turns out that at least some individuals are. Individuals at least sometimes have observational awareness of the extent to which their own actions are biased; and sometimes they additionally have what we can refer to as attributional awareness: awareness that some feature of their cognition – implicit prejudice – can be attributed as causally producing that influence.7,8

We have three candidate senses of awareness, then:

(a) introspective awareness of the implicit association itself, or its operation;
(b) inferential awareness of the body of knowledge about people's tendencies to harbour, and display, implicit bias;
(c) observational awareness of the effects of the implicit associations on behaviour (sometimes alongside attributional awareness of the cause of these effects).

Because these distinct senses of awareness have not been distinguished in either the empirical or philosophical literatures, we should be cautious in our claims about the relationship between these different senses of awareness; considerably more conceptual (and empirical) work is needed to understand their relationship than is possible here. However, some preliminary remarks can be made. Firstly, if one denies that introspective awareness of implicit biases (or indeed any mental state, cf. Levy, 2014) is possible, then acquiring knowledge of the other kinds (inferential or observational) will not garner that sort of awareness. Secondly, gaining inferential awareness (of the fact one is likely to be biased) may well aid observational awareness, if it can prompt reflection on one's behaviours that might yield evidence of subtle discrimination. Finally, inferential and observational awareness of facts about implicit bias may also generate attributional awareness, in that one may be able to attribute one's discriminatory behaviour to the probable presence of implicit bias, even if one is unable to introspect on such biased processes.

2.6. A normative epistemic condition

Each of these senses of awareness of implicit bias has been used in the literature regarding whether individuals meet the epistemic conditions for responsibility. How should we think of these different senses in relation to the epistemic conditions for responsibility?

When considering whether lack of awareness, or ignorance, exculpates, we have to ask not merely what an individual does or does not have awareness of, in some relevant sense of awareness.9 Some failures of awareness are themselves

---

7 We can think of attributional awareness as a subset of inferential awareness. Crucially, though, it is not concerned with an inference from knowledge of general empirical evidence of implicit bias to one's own states; these individuals had not been given information about the evidence from empirical psychology. Rather, it is an inference from an observation about one's behaviour to a feature of one's cognition.

8 This seems to be the sense of awareness that Levy is concerned with where he describes the 'indirect route' by which we come to have knowledge of our implicit attitudes, namely, via self-interpretation. This is not entirely clear, though, because Levy does not distinguish between different kinds of awareness, and at some points talks of awareness of a 'gut instinct' which is not obviously an implicit association. Moreover, he is working within a framework that requires that we assume (for the sake of argument with King & Carruthers, 2012) that introspective access to our attitudes (implicit or explicit) is not possible. In contrast, I am not ruling out the possibility of introspective access to some attitudes, and hence the contrast between introspection of some attitudes, and observational awareness of implicit attitudes, is available here. Levy's argument against moral responsibility rests not on awareness – he acknowledges that individuals may have (observational?) awareness of their implicit attitudes – but on the structure of implicit attitudes and the functional role they are unable to play in our moral agency. I address in detail his arguments in Holroyd & Kelly, 'Implicit Bias, Character and Control' (in press), arguing that we should reject his claim that implicit attitudes cannot play the relevant role in our agency.

9 This is the 'spotlight' view, whereby only that which falls under the spotlight of conscious awareness is a candidate for what the agent can be held responsible for. But Sher effectively undermines the spotlight view, showing that whilst it is often asserted in philosophical contexts, it does not cohere with our common sense or basic intuitions about who is responsible for what; nor is it philosophically defensible.
culpable. For example, ignorance of the harm one is perpetrating may not exculpate if one can reasonably be expected to check that one’s actions are not harmful in that way. Similarly, forgetting – being unaware of – a meeting one had promised to keep does not exculpate, and may itself be culpable, in the usual course of things (setting aside excessive pressures or stresses or distractions). Not being aware of a motive of cruelty or jealousy that shapes one’s interactions with friends does not excuse actions that are cruel or express jealousy. This is because, in all these cases (bar exceptional circumstances) we think that the agent should be aware of the morally relevant facts – whether they are causing harm, or are forgetting a commitment, or expressing a cruel motive – and their failures of awareness or knowledge are themselves culpable. (See Sher (2009) for a recent articulation and defence of the epistemic conditions for responsibility in these terms. The reasons for this culpability need further unpacking; something we return to in Section 4.).

So rather, we should proceed by asking whether individuals are in fact aware, in the (to be determined) relevant sense, of their implicitly biased actions; but rather, whether an individual should be aware in this sense, and whether their failures of awareness are culpable.10 We are asking what individuals should know qua responsible agents.11 This approach is consistent with the idea, defended in legal philosophy, that negligence does not require that an individual in fact be aware of the harm caused by her action; only that a reasonable person would have been.12 So our question is now more well focused: do the epistemic requirements that apply to responsible individuals include awareness of the kinds identified above? In the next section, I consider each requirement in turn in order to identify the relevant epistemic conditions for responsibility for implicitly biased actions.

3. Epistemic conditions and responsibility for implicit bias

We are interested, then, in whether the following claims are true as conditions for responsibility:

(a) individuals should be aware, introspectively, of the operation of implicit associations
(b) individuals should be aware of, or know about, the body of knowledge about people’s tendencies to harbour and display implicit biases, and make the relevant inferences about their own tendencies to express implicit biases.
(c) individuals should have observational awareness, or knowledge, of the effects of implicit associations on their behaviour.

How might we proceed in evaluating which kinds of knowledge individuals should have? There is a methodological difficulty here, in that we have quite a few moving parts. The conditions for moral responsibility are in question, but the mental phenomena, processes, and actions they influence are in some respects unfamiliar – they are states about which we are learning ever more from the findings of empirical psychology. What should be held fixed, in trying to understand how we should think about the role of these unfamiliar processes in our agency?

In question is not whether we are ever morally responsible for anything; that we are, at least sometimes, is a starting assumption of the argument.13 Those who argue that we are not responsible for actions influenced by implicit biases (or for the implicit associations themselves) are not seeking to vindicate general scepticism about responsible agency. At issue, rather, is whether certain aspects of our agency fall into the remit of those things for which we are responsible. One strategy we can employ is to consider other more familiar aspects of agency, and our judgements about responsibility regarding these. Should the same be said of implicit biases? Another strategy is to consider where a certain condition for moral responsibility would set the bar if it ruled out holding individuals responsible for a certain kind of state or behaviour: would it set the bar implausibly high, and lead us to general scepticism about responsible agency? Should certain aspects of our agency fall into the remit of those things for which we are responsible? One strategy we can employ is to consider other more familiar aspects of agency, and our judgements about responsibility regarding these. Should the same be said of implicit biases? Another strategy is to consider where a certain condition for moral responsibility would set the bar if it ruled out holding individuals responsible for a certain kind of state or behaviour: would it set the bar implausibly high, and lead us to general scepticism about responsibility? I will deploy each of these strategies in proceeding.

In asking whether we should consider each epistemic condition as specifying a requirement for responsibility, we should ask whether it is a) desirable, and if so, b) possible, to meet the norm. For if it is impossible to gain awareness of some sort, then it would be unreasonable to require that individuals have such knowledge as a condition of responsibility. So our evaluation will require attention both to philosophical questions about the defensibility of certain conditions, and empirical questions about what sorts of awareness are possible, as far as the findings of empirical psychology reveal.

3.1. Individuals should be aware, introspectively, of the operation of implicit associations

Is this a requirement on responsible agency that we should endorse? If so, then the non-culpable failure of individuals to have introspective awareness of the activation and operation in their cognitions of implicit biases would indeed mitigate

---

10 Washington & Kelly frame the issue in terms of ignorance exculpating, except where the individual is responsible for now knowing X. This seems to me to suppose the default view that ignorance exculpates, but I am not entirely happy with that as a starting assumption, so prefer to frame the issue in terms of what individuals should know, and whether failures to know what they should know are culpable.

11 I am treating the issue of what individuals should know, and what they should be aware of, interchangeably. But there might be some cases where individuals have awareness that amounts only to true beliefs (say, if that awareness lacks the justificatory grounds to constitute knowledge). Thanks to Kengo Miyazono for emphasising this point.

12 See e.g. Ashworth & Horder 2013, p. 182.

13 Some theorists have supposed that the findings of empirical psychology give us reason to revise our conceptions of agency and responsibility, with the consequence that these notions are radically different from both common sense and earlier philosophical formulations (see e.g. Bargh, 2005; Doris, 2002). These conclusions seem to me not to be supported, and I do not think we are pushed to a radical scepticism about agency or responsibility.
from responsibility. My inability to detect the activation of an implicit association, or the extent to which it is influencing my cognitive processes, would suffice to absolve me from responsibility for any influence that association exerts on behavioural outputs.

It should be noted that any failure of this kind would, seemingly, be non-culpable. Many have claimed that the operation of implicit associations – their activation and role in our cognitive processes – is not something to which we have access simply by reflecting, or by introspectively checking (see e.g. Saul, Washington & Kelly). And these assertions are borne out by empirical studies also. Indeed, various studies rely on the opacity of various aspects of our cognition. For example, the ‘evaluative priming test’ operates by presenting a prime (e.g. a black or white face, a picture of a prominent Republican or Democrat) and then measuring how long it takes individuals to recognise and categorise a negative or positive word. The idea is that if an individual is faster to categorise the negative terms as bad, than positive words as good, this reveals a negative association with the prime (because negative constructs were made more accessible by the prime). The idea is that individuals do not have introspective access to the ways in which the prime (of which they may or may not be aware) activates certain associations which influence their ease of categorisation. Such implicit processes in general do not appear to be ‘operational transparency’ to us, such that it is not possible to have introspective awareness of their operation. If so then any failure to have such introspective awareness would be non-culpable.

But should we endorse this requirement as a condition for responsible agency, such that non-culpable failure to meet it does exculpate? This requirement does not seem to me to be defensible. Firstly, it is not a standard we apply to other aspects of our cognition. Secondly, were such a standard to be applied, it would lead to radical scepticism about the possibility of responsible agency.

On the first point: we should note that with respect to other aspects of cognition, it is not a condition on responsibility that individuals be aware of the cognitive processes that produce action. Here are two examples that help us to see this. Consider a case discussed by Nancy Snow (2006), of an individual instinctively making an intervention when she observes an elderly woman being cheated by a sales clerk (556–557). A central feature of Snow’s example is that the agent does not recognise that her sense of justice is activated (her justice related goals, in Snow’s terms); there are important aspects of her cognitions, then, in relation to which she lacks introspective awareness. But that she lacks awareness of this aspect of her cognition does not mean that she cannot be held responsible and praised for her actions.

Likewise with blameworthy actions. Cases of forgetting are clear candidates of instances in which individuals might be blamed, despite lack of awareness of whatever processes led them to forget (indeed, awareness of this might allay the forgetting!). Consider Sher’s discussion of forgetting for which the agent is responsible: a distracted parent forgets that a dog is languishing in an overheating car. The parent lacks awareness of the cognitive processes whereby various competing demands crowd out the relevant belief; this leads her to act in a negligent and harmful way. But Sher asks us to share the intuition that the agent is still responsible for this failure, despite the lack of awareness of the processes (the failures of attention) that produced the action.

Here are two cases, then, in which it is plausible that we hold an agent responsible (for creditable action, and for blameworthy action) even whilst they fail to be aware of the cognitive processes that play a role in producing the actions for which we hold them accountable. Thus, it seems that we do not require introspective awareness of the processes involved in the production of action as a condition on responsibility. My contention is not, of course, that these examples are exactly the same as cases of implicitly biased actions. There might ultimately be different judgements to be made about the two kinds of cases. However, what these cases show is that merely lacking introspective awareness of the processes involved in deliberation and action does not suffice to exculpate, and is consistent with praiseworthy and blameworthy action. Nonetheless, the more familiar processes described above are similar in some important respects (whilst of course dissimilar in others) to those involving implicit associations that produce implicitly biased actions: they are fast, automatic, not readily under the agent’s deliberative control, unreflective, and (in the latter case) processes the outputs.

Likewise, were such a standard to be applied, it would lead to radical scepticism about the possibility of responsible agency.

14 For discussion of the EPT as an implicit measure, see Fazio Petty and Briñol (2009) and Bar-Anan and Nosek (in preparation).

15 For in general, we do not maintain that individuals should have this kind of knowledge of their cognitive processes; we do not require ‘operational transparency’ for individuals to be held responsible for actions that result from these processes. Indeed, if we did, very many of our actions – perhaps all of them? – would be exempt from responsibility, insofar as we are never aware of all of the processes that input into the production of action. Lack of awareness of this kind, then, does not excuse responsibility. Even if it is not possible to secure this kind of awareness, that is irrelevant: because introspective awareness is not a requirement for morally responsible action.
3.2. Individuals should be aware of, or know about, the body of knowledge about people's tendencies to harbour and display implicit biases, and make the relevant inferences about their own tendencies to express implicit biases

Let us turn, then, to the second formulation of the epistemic condition for responsibility in relation to implicit biases: that individuals should be aware of the relevant facts about the tendencies to express implicit biases in action. This is the sort of condition that Washington & Kelly argue for, in developing their ‘externalist’ account of the epistemic conditions for responsibility.

The condition on responsibility here is one which maintains that individuals should, in some contexts, have an awareness of the facts uncovered by empirical studies about tendencies to discriminate as a result of implicit bias; inferences can then be made about one's own propensity to do so. (Ultimately, steps to avoid biased actions can then be taken). There are two distinctive moves made by Washington & Kelly in advancing this condition: first, Washington & Kelly argue that what individuals should know (with respect to bodies of knowledge such as that about implicit bias) is role dependent in a significant sense. For example, a football commentator is required to keep up with changes to team composition in the seasonal transfer market; a heart surgeon is required to have up to date knowledge of aortic valve technologies; I am not required to have knowledge of either. The second key move is to point out that the possibility of meeting this requirement is dependent upon one's epistemic environment. Accordingly, individuals could not meet the requirement to know about implicit biases in 1980, when the findings of empirical psychology about implicit cognition were less well established and less readily available. In this way the epistemic conditions for responsibility are ‘externalist’ in a significant sense – depending not on facts about the agent, but on facts about her epistemic environment – what is known, and what is available to be known.

This leads Washington & Kelly to conclude that we should endorse this condition for moral responsibility, (b), in the case of individuals who (i) need to have information about implicit biases because of the social role they occupy; and (ii) have that information available in their epistemic environment, such that they are responsible for not availing themselves of that knowledge. Thus they maintain that responsibility ‘accrues first, or at least more quickly and disproportionately, to occupiers of specific social roles. These include those involved in hiring decisions, obviously, but also teachers, social workers, and those in other “gate keeper” positions whose activities can have the most amplified effects on various institutions and population level outcomes’ (in press). This issue speaks to the extent to which failure to meet this epistemic condition is culpable – is such ignorance itself a culpable failure to fulfil one's epistemic responsibilities? For the contemporaneous hiring committee member, according to Washington & Kelly, it is: they can be held responsible (and perhaps blamed) for not knowing what they should – given the availability of the relevant information, and the role they occupy – be aware of. An implication of this is that for other individuals – the proverbial ‘person on the street’ – the lack of awareness of this knowledge is not culpable; her failure to meet this epistemic requirement does not mean she fails to grasp something she should grasp – so she, unlike the ‘gatekeepers’, can be exculpated from responsibility. This, Washington and Kelly maintain, is due to the fact that our epistemic environment is one in which knowledge about implicit bias ‘has still not risen to the level of common knowledge, and it probably will not any time in the immediate future. . . and the percentage of people who have not heard of implicit bias at all is probably still quite high’ (in press). Accordingly, the extent of this exculpation depends on the ‘availability’ in the epistemic environment of the relevant knowledge. When it is pervasively known, and not just within the remit of academic researchers, this failure will be more widely culpable.  

Should we agree with Washington & Kelly's statement of this epistemic condition for responsibility? To start, we should ask ourselves what motivates the move to restrict the normative requirement (should know) to those in 'gatekeeper' roles. Why should such individuals know about implicit biases? The answer is presumably that individuals in such roles are more likely to manifest implicit biases in ways that we can reasonably foresee to have deleterious effects on those who are stigmatised by or discriminated against by those biases, or in ways that affect population level distributions of benefits and burdens. But this motivating assumption seems to me implausible. Many people make decisions about who to hire, who to fire – literal ‘gatekeeper’ decisions – but also about who to grant a loan to, where to live, who to stop and search, who to give a lift to, what news stories to report (and how), who to write prescriptions for, who to sit by on a train, how to evaluate co-workers, who to smile at, what grades to assign or references to write, who to cross the road to avoid, who to believe, who to befriend . . . and so on. These kinds of interactions can all be affected by implicit biases (see Jost et al., 2009 for an overview). And it seems to me to be difficult to substantiate the claim that the reasonably foreseeable cumulative effects of these interpersonal interactions are not greater than those of the gatekeepers who make decisions with population level effects. The widespread impact of ‘informal’ discrimination and segregation in sustaining patterns of disadvantage is described in detail in Anderson's work on racial inequality (2010), Valian (1999) also emphasises that population level discrepancies in the distribution of advantages can often be traced to the accumulation of small instances of differential treatment.

Accordingly, it seems to me implausible to restrict the realm of ‘responsibility to know’ to some few individuals charged with hiring decisions or other population level outcomes, especially given what we know about the pervasiveness of implicit bias and its effects. If this is right, then the epistemic requirement to be familiar with the findings of empirical psychology

---

16 See also Vargas (in preparation) for discussion of the idea that our ‘moral ecology’ does not currently support holding individuals responsible for bias, but that we should work towards creating the conditions in which this is the case. Vargas’ argument turns on whether our ‘moral-considerations sensitive’ capacities are appropriately sensitive to facts about implicit bias, and whether the efficacy of our practices of holding each other responsible – in terms of helping us become better moral agents – is promoted or undercut by currently holding each other responsible for implicitly biased actions.

17 For a discussion of the moral issues surrounding the role of implicit biases in our personal relationships, see Sheila Linott (in preparation).
will apply very widely indeed. Almost everyone will be subject to the requirement that they are aware of the relevant body of empirical findings about the nature of implicit biases, which of their actions are likely to be susceptible to it (and ultimately, how to avoid this). This knowledge is available in their epistemic environment, and it is knowledge that they should acquire. This seems like a very demanding epistemic condition!

Washington & Kelly might appeal to the idea that such knowledge, whilst available, is not readily accessible as ‘common knowledge’, so failure to grasp it is not culpable. But the general epistemic requirement cannot be that the knowledge is ‘common’ in this way – otherwise the hiring committee members would also fail to be culpable (these gatekeepers, too, are making decisions where knowledge of implicit bias is not common knowledge). Moreover, if the potential and cumulative damage foreseeable as a result of the biased actions of, for example, loan-makers (and prescription-writers, testimony-takers, stop-and-searchers, and so on) is commensurate with that of hiring committee ‘gatekeepers’, there is no reason to subject them to different epistemic standards.

But if the same epistemic standards apply more broadly it looks like we are committed to maintaining that very many individuals in fact are responsible for acting in an implicitly biased way, because they are culpably blameworthy for not knowing what they should know (namely, about a body of knowledge in empirical psychology). This seems rather implausible: not because it involves maintaining that very many individuals might be responsible for acting in implicitly biased ways (I think we may be); but that the grounds for this are the failures to engage with the findings of empirical psychology.

To summarise: plausibly, very many of us play a role in sustaining and perpetuating patterns of disadvantage by acting in ways that are implicitly biased; also plausibly we have a responsibility to avoid doing so. But Washington & Kelly seem to be committed either to denying this, or to the claim that, therefore, almost everyone has a responsibility to engage with the findings of academic research. This is an implausibly demanding epistemic requirement, and one that accordingly, we should reject.

One of the reasons we should reject this condition is because it supposes that the only access individuals might have to knowledge about implicit bias is via (some perhaps mediated) academic research. But there is reason to believe that this is not the case. This brings us to the third epistemic condition for moral responsibility.

3.3. Individuals should have observational awareness, or knowledge, of the ways their behaviours are influenced by implicit associations

Let us now turn, then, to the third sense of awareness identified earlier. I have suggested that individuals cannot, and ought not to, have awareness of all of their (or the relevant) cognitive processes prior to acting; and that it is unreasonable to demand of almost everyone that they engage with complex bodies of academic knowledge about implicit biases. But there remains the question of whether individuals can, and should be expected to, have knowledge of the morally relevant features of the actions they perform. In this context, the morally important features include the property of being discriminatory, or treating differentially on the basis of some arbitrary feature (such as race, gender, or age).

Do we in general require that individuals have awareness of these properties as a condition of responsibility? Do we require that individuals should have this knowledge? And to what extent is failure to have that knowledge culpable? Let us again consider two cases that will help us to think about this condition in general, before we turn to consider it in relation to our concern with implicit bias.

Let us consider a variant of a case presented by Adams (1985), in which an individual acts in a way that expresses ingratitude; perhaps she is insufficiently warm in accepting a favour, say. This individual has not confronted the fact that she harbours this attitude of ingratitude, Adams tells us. And this is because she cares more about having a good opinion of herself than confronting unsavoury truths about herself. In being unaware of the presence of this attitude in her cognitions, the agent will therefore be unaware of her action manifesting this attitude; so unaware that her behaviours manifest the morally undesirable characteristic of being ungrateful. But we should nonetheless hold the agent responsible for the actions that are inflected by this attitude, Adams claims, because she should be aware of the fact that her actions are inflected by this attitude.

If we share Adams’ judgements, then we should hold that it is not a condition for responsibility that the agent is in fact aware of the morally relevant features of their behaviour. A failure to know that one’s action is expressive of ingratitude does

---

18 An anonymous reviewer has suggested that we might also appeal to the idea that hiring committee members have explicitly taken on such a role. But this voluntarist picture of our epistemic obligations does not sit well with Washington & Kelly’s claim that, when information becomes common knowledge, individuals should then grasp it, and will be culpable if they do not (irrespective of whether they have voluntarily undertaken a role that requires it).

19 Note that my claim is not that individuals are not responsible for engaging with the deliverances of science. Rather, it is not reasonable to expect every member of society to do so when it is infeasibly costly (because, e.g. research published in academic journals is behind paywalls). I am happy to grant Washinton & Kelly the claim that knowledge about implicit bias is not yet common knowledge, though it also seems right to note that we are in a transitional phase of moving towards such an epistemic environment, via diversity training that includes information about implicit biases and popular non-fiction books that make the phenomenon more widely acknowledged. Thanks to an anonymous reviewer for pressing that point.

20 An alternative option is to say that it is the responsibility of academic researchers to make available this information in accessible formats so as to enable almost everyone to become familiar with it. This would make the epistemic requirement less demanding for the proverbial person on the street. We could accept this, but – as I argue below – there are in any case other defensible conditions for responsibility, and I believe these better explain why, even in the absence of such measures, individuals may be culpable for implicitly biased actions.

21 This is not to say that – especially in some cases, such as those of hiring committee members – is it not desirable for individuals to acquaint themselves with this information. Rather, it is to say that having such information is not necessary for moral responsibility, and it is implausible to demand that very many individuals should have this sort of knowledge.
not absolve one of responsibility for acting ungratefully. This is something the agent should be aware of. In Adams’ version of the story, the agent’s failure of awareness is due to some further agential fault (the desire to have a good opinion of herself). Moreover, in this case, it seems clear that this is something that, with a bit of reflective work, the agent could become aware of. Observing that her action manifests ingratitude, then, is something we can reasonably expect the agent to do.

Let us consider a second case, this time from Sher (2009). Sher presents an individual who tells an anecdote, failing to notice that in doing so she is being insensitive to her audience (an anecdote about a financial failure that is not well received by the individual who has recently experienced financial failure, say). Her behaviour has the property of insensitivity, but this is something of which she is not, at the time of her remarks, aware. Nonetheless, Sher urges us to share his intuition that this individual is responsible and blameworthy for her insensitivity. This is because she should be aware of her insensitivity, he claims, even if she is not.22

If we agree with Sher, then, again, we should not hold that it is a condition for responsibility that the agent is in fact aware of the morally relevant features of her behaviour (in this case, insensitivity). The agent should be aware of this feature, and we can suppose that this is something that, with sufficient reflection, it would have been possible for her to notice about her behaviour. Given this, it is a reasonable expectation to hold the agent to. These cases seem far from controversial, and are instances of common parts of our practice of holding each other responsible. So far, then, the epistemic condition I have described in this section is far from revisionary.

Finally, then, let us consider whether the same considerations apply to the case of actions that express implicit bias – such as the differential evaluation of CV whose only difference is the racialised name at the top. These actions have the morally undesirable characterising of being discriminatory. If we treat this case as a direct analogue to those considered above, then the mere fact that an agent is not aware that her action is discriminatory will not suffice to exculpate. Ignorance alone does not excuse. So we turn to our next question: is awareness of the discriminatory nature of her action something that we think the agent should have knowledge of? Certainly we should strongly affirm the desirability of this: insofar as agents are expected to bring their behaviour into conformity with moral norms, we should insist that they ought to self-monitor, and be aware of ways in which their behaviour may depart from these moral standards. But whether this expectation is reasonable in the case of implicit bias depends on whether it is possible for agents to have this awareness. Given the implicit nature of the biases at issue, shouldn’t we suppose that this is not something that individuals can reasonably be expected to be aware of?

As we have already seen (1.5 above), the empirical evidence does not provide support for that claim, and rather indicates the contrary. First, it is worth noting that whilst some authors have characterised implicit associations as attitudes of which the agent is unaware (Saul, Washington & Kelly, see above), DeHouwer et al. (2009, 357–358) point out that there is in fact little reason to endorse this claim. Whilst in some implicit measures the way in which the association is activated is something that the agent is not aware of (because the association is primed) this does not mean that the effects of the bias on behaviour are not something that the agent can be aware of. Likewise, that an attitude is measured with an implicit measure (a measure that does not require self-report or reflective articulation of one’s attitude) does not mean that the behavioural manifestation of the attitude cannot be reported on.

Recall the studies by Monteith et al. (2001), which I mentioned earlier in introducing the notion of observational awareness. In this study, participants undertook a race IAT (pairing white or black names with pleasant or unpleasant terms). The participants were then asked to evaluate their performance on the IAT, and the interesting finding for present purposes was that a significant number of participants (64%) were able to report on the basis of observational awareness of their own responses that they responded more slowly when pairing black names with pleasant terms (than with unpleasant terms, and than white names with pleasant terms).

These findings garner support from more recent experimental studies.23 Hahn, Judd, Hirsh, and Blair (2013) examined individuals’ accuracy of predictions regarding the expression of implicit biases. They found that individuals were, when asked to carefully reflect, able to accurately predict this, both in experimental terms: ‘My sorting of [the congruent pairings] will be very/moderately/slightly easier...’ (p.5) and in conceptual terms: ‘My true implicit attitude is a lot/moderately/slightly more positive towards white’ (p.8). This was so even where individuals showed discrepancies between recorded implicit attitudes, and reported explicit attitudes, such that predictions were not being made on the basis of explicit attitudes of which the participants were aware and alert to the possibility of their subtle influence on behaviour.24

We might think that the participants made their predictions on the basis of general knowledge of social context rather than on the basis of awareness of their own behavioural dispositions. But the experimenters ruled this out by asking participants to predict both their own, and the average responses. There was divergence between these predictions, with predictions about their own behaviour more closely matching biases measured (p. 10). Or, we might think that reports of explicit

---

22 A crucial further feature of Sher’s account is that the agent’s failure is traceable to psychological structures that underpin her agency, such that the omission in question (the failure to know what she should know) is suitably attributable to her. These remarks seem to me to be defensible, and cohere with the suggestive proposals I make in the final section of the paper. I have also remarked on the ways in which at least some implicit attitudes are bound up with agents’ values in my 2012, and in Holroyd and Sweetman (in press).  
23 Levy cites Ranganath, Smith, and Nosek (2008) as providing evidence that individuals have awareness of their implicit attitudes (2014, pp.10–11). However, this study indicates (as Levy acknowledges) that individuals are able to report on their ‘gut instincts’ which were found to correlate with implicit attitudes. But these gut instincts may not be the same as implicit attitudes, and so it does not seem to me that Levy’s inference from this finding to the claim that individuals are aware of their implicit attitudes is supported by this evidence.  
24 It is important to note that these results were found both individuals who had considerable implicit association test training and experience, and those who did not. So the predictions do not seem to be based on familiarity with the tests and what it shows.
attitudes are unreliable, such that individuals in fact had explicitly biased attitudes, and made reports in anticipation of these attitudes influencing their behaviour. This possibility seems unlikely, given that the effect was found in the condition in which participants were told that the results did not reveal their ‘true attitudes’, but merely cultural associations (manipulation checks revealed that these participants accepted this story) (p.7).

This study is important, because it indicates that individuals are not only able to detect morally relevant features of their actions post hoc; they were also able to predict morally undesirable features ex ante. In a way, this should not seem surprising given the remarks above: there is nothing about the manifestation of an implicit association or attitude in behaviour that prevents it from being accessible to report on. But it is surprising in the context of philosophical discussions that have supposed that implicitly biased behaviour is something of which individuals are not (in some sense) aware, and have elided the different notions of awareness at issue. But these assumptions are not supported: there is evidence that supports the claims that, with reflection, individuals are at least sometimes able to detect and predict discrepant responses. These are tentative findings: we might wonder whether the possibility of detecting and predicting biased responses extends to the full range of behaviours that might be influenced by bias. This is particularly so where the biased actions are identified in studies that observe statistical tendencies across groups, rather than in intrapersonal differences in responses. And we should want to know more about the functioning of this awareness ‘outside the lab’. Nonetheless, the present findings provide reason to examine how widespread observational awareness of one’s actions as biased is, given that it is at least possible sometimes to have such awareness.

The key points, then, are as follows: firstly, individuals should have knowledge of the morally relevant features of their actions, such as that they are discriminatory. It is only if individuals non-culpably lack knowledge of this sort, then, that they should be absolved of responsibility for discriminatory, implicitly biased behaviour. One way of non-culpably lacking this knowledge would be if it were not possible to have such awareness. But the evidence suggests that this sort of awareness is not ruled out; and moreover, sometimes individuals are able to gain this sort of awareness. This raises the question, then, of whether there are other grounds for supposing that individuals who lack awareness of the ways in which their actions are inflected with bias are culpable for this lack. I suggest two possible (non-exhaustive, non-exclusive) explanations that might implicate other imperfect cognitions in our responsibility for implicitly biased behaviour.

4. Implicit bias and other imperfect cognitions

We have arrived at the following question: when individuals lack the awareness that it appears possible for them to have, with respect to the morally relevant properties of their implicitly biased behaviour, is this lack a culpable one? Not all failures to know what should be known are culpable failures. What might be said about the failure to have the relevant kind of awareness of one’s own actions as implicitly biased? Here are two possible answers that yield different judgements about the extent of an individual’s blameworthiness for their actions.

4.1. Failures of attentiveness

It might be tempting to suppose that the kind of awareness at issue requires some revisionary moral understanding, such that the culpability of any failure to meet the epistemic condition is considerably mitigated. Not knowing what one should know may be less culpable if it is harder to gain that knowledge, because it is not yet normalised and nor accessible to all as moral knowledge (it is rather at a ‘frontier’ of moral knowledge) (Calhoun, 1989, cf. Washington & Kelly, above); or if the pervasive tendency is not even to think of an issue as ‘morally charged’, so that moral reflection is not focused upon that action or its consequences at all (Isaacs, 1997). Examples of this include the failure of many to know that the supposedly gender-neutral use of the pronoun ‘he’ perpetuates sexism. Such sexist language use may be less blameworthy if the knowledge of this that we all should have is not yet normalised, or requires some imaginative leap to access; or if it is not yet clear that this domain of our activity is even morally scrutable.

Analogously, we might think that the common sense view of discrimination involves the explicit intention to treat differently, or the explicit intention to harm others or manifest ill will (Garcia, 1996). Given this, the failures of awareness with respect to implicitly biased actions may be candidates for a kind of moral ignorance that is mitigated in the ways described above. Perhaps the knowledge that one’s actions can be discriminatory, even if not intentionally so, is not yet ‘normalised’ so

[25] See also Gawronski, Hofmann, and Wilbur (2006) for further discussion of whether implicit attitudes are unconscious.

[26] The problem of false negatives should be noted: namely, that in those who did not detect or predict a discrepancy, there was nonetheless some implicit bias measured. So, whilst detecting implicit bias is possible, supposedly discerning that one is not implicitly biased is not reliable. This asymmetry raises interesting questions that I address in 3b.

[27] Should we endorse another epistemic condition, namely, that individuals are aware of how they might change their associations or behaviour? We might endorse this if we thought that a necessary condition for responsibility were that individuals could do otherwise; then this awareness would be important for the fulfilment of that condition. But I do not endorse that condition. In contrast, the reason for holding that it must be possible for the agent to know what she should know, is that it ties responsibility to epistemic states that are in principle available to the agent, but of which occurrent awareness is (culpably or non-culpably) lacking due to other aspects of her psychological make up. This is not to say that having such awareness – and introspective or inferential awareness – would not be morally desirable; only that they are not necessary conditions for responsibility for implicitly biased actions. Thanks to James Andow and Chloe Fitzgerald for pressing me on these issues.

[28] See also Fricker, 2010 for discussion of the different demands of our practices of holding each other responsible across different moral contexts.
as to make this knowledge readily accessible; or perhaps such behaviour does not prompt reflection, as it is not on the radar as a ‘moral issue’ at all. In this respect we can see how inferential awareness of the findings about implicit bias may make it considerably easier to possess observational awareness of one’s actions as biased. Knowing that one is likely to be biased, given the body of empirical evidence we now have, can introduce pressures to reflect that might generate observational awareness. To the extent that one lacks inferential awareness, then, it may be harder to know what one should know. Moreover, we might think culpability is mitigated further if this moral ignorance is compounded by misleading introspective evidence, which seems to suggest to us that our motives are good, and without discriminatory content.

However, these remarks do not seem quite right as a full diagnosis of the failing, because what is at issue is not simply whether a certain behaviour is morally problematic or would fall under the rubric of ‘discrimination’; but rather whether the fact that a behaviour displays differential treatment is noted at all. The lack of awareness at issue here is not whether the situation is a ‘morally charged’ one; it is surely accepted that one’s behaviour would be morally questionable if it were known to involve such differential treatment. Rather, the lack of awareness pertains to the behaviour manifesting differential treatment itself. There is some lack of attentiveness, a failure to notice, that means the morally relevant feature of behaviour is not something of which the agent is aware (irrespective of how it would, once noticed, be labelled).

It might be that this failure of attention is driven by the mistaken beliefs that, since one is not intentionally discriminatory, it is simply not possible for one’s actions to have discriminatory effects. But on the other hand, some failures to notice morally important things can be indicative or expressive of an agent’s evaluative stance – what they care about, and how much. Lacking the motive to reflect can indicate what one takes to be worthy of moral scrutiny. Such lacks are ones that it is feasible to hold individuals responsible, and blameworthy, for (Sher, 2009; Smith, 2005). The extent to which an individual is culpable for the failures of awareness with respect to implicitly biased actions, then, will depend on answers to further questions about the role of implicit associations in our broader agential structures, and how their expression is related to the values we hold.

4.2. Self-deception

It is one thing to lack a motive to reflect; another to be motivated not to reflect, as Calhoun points out: ‘self-interest can motivate the suppression of reflection ... self-deception is a matter of not being motivated to examine one’s actions or reasoning too carefully, lest something unpleasant turn up’ (399). Whilst it might be possible to detect in one’s actions differential treatment – or to predict it – when specifically prompted to do so, these are difficult truths to confront. Acknowledging not only one’s complicity in but perpetuation of patterns of discrimination, albeit in subtle and unintended ways, is something that no doubt many of us find hard to accept. The belief that one’s actions are implicitly biased, and other implied beliefs about one’s role in sustaining patterns of discrimination, are clearly beliefs that, for a range of reasons, agents might be motivated not to confront. Conversely, the belief that one’s actions are consistent with one’s moral ideals (of non-discrimination, of being evidence sensitive and unbiased) is one that agents are motivated to maintain.

That we are motivated to avoid the sort of moral reflection that might overturn those desirable beliefs, and turn up these undesirable facts about our propensity to bias, is a plausible explanation of this lack of awareness. After all, there is ample evidence from empirical psychology that we are motivated to maintain a positive self-concept (Brown, 1986; Suls, Lemos, & Stewart, 2002). Indeed the motive to present a positive view of ourselves is what raises concerns about the validity of self-report measures (whether we want to see ourselves, or let others see us, as prejudiced), and makes access to attitudes via implicit measures so valuable. Moreover, recent studies indicate that such a motivation has a role in sustaining our view of ourselves as immune to bias: Pronin and Kugler (2007) found individuals to over-identify on misleading introspective evidence of propensity to bias (introspection revealing – surprise! – no bias). Meanwhile participants ignored behavioural evidence of their own bias that they were willing to take as evidence of bias in others: ‘actors . . . preferred to see themselves as bias free’ where it was possible to ignore evidence to the contrary (576).

This form of self-deception could explain the failure of individuals to have awareness that their actions manifest implicit bias. Would this explanation yield the judgement that such ignorance is culpable? Whilst pervasive, such self-deceptions are not so overwhelming as to be insurmountable: in Pronin and Kugler’s study, when experimenters reminded subjects of the unreliability of introspection (in contrast to observed behaviour) as a guide to their own bias, the participants no longer denied their susceptibility to bias. Insofar as the motivated lack of reflection is serving to bolster a misleadingly positive view of oneself, and serving to cover up morally undesirable aspects of one’s actions that are not otherwise impossible to detect, a case can be made that failures of awareness resulting from such mechanisms are culpable. In this case, such ignorance would be culpable, and the lack of awareness of one’s actions as implicitly biased would not serve to exculpate from responsibility.

29 I make some preliminary remarks about how some implicit associations appear to be bound up with our values in my 2012.
30 The sort of self-deception I have in mind here, then, is that captured by non-intentionalist accounts of self-deception (which do not require an intention to deceive oneself, but only that one has motivated beliefs). See e.g. Mele, 2001; Barnes, 1997.
31 See Nosek, 2005, 2007.
32 An alternative (and not necessarily competing) explanation could appeal to the confabulatory explanations individuals give for their actions, filling in some ‘gap’ in the justification for their behaviour. See Ema Sullivan-Bissett (2015) for discussion of the nature of confabulations, and their role in explaining biased behaviour.
5. Summary

I have argued that there are three different claims that need teasing apart when thinking about whether individuals have awareness in relation to implicit bias: whether we have introspective awareness of its operation in influencing behaviour; whether we are aware of the relevant bodies of knowledge about implicit bias that enable us to infer our likely susceptibility to such biases; whether we are observationally aware of the fact that our behaviours have the morally undesirable property of being discriminatory. In each case, the relevant question is not whether an individual has this awareness, but whether they should have such awareness, and whether lacking it is culpable. I argued that lacking introspective awareness of the operation of implicit associations, or lacking inferential awareness of the propensity to display implicit bias, does not in itself exculpate. Rather, I argued that we should have observational awareness of the morally relevant features of our behaviour, namely, their discriminatory nature. And indeed the empirical studies on implicit bias do not entail that such knowledge is impossible to gain; in fact, some show that sometimes individuals do have this kind of awareness. When we lack it, are we culpably ignorant? Perhaps so, if this lack is due to failures of attention that express our values, or self-deception that narrowly serves our interests. In this way, our responsibility for implicitly biased actions may be bound up with imperfect cognitions of other kinds.

Acknowledgments

This paper has benefited from discussion with and feedback from the participants in the Epistemic Costs and Benefits of Imperfect Cognitions Workshop, organised by Lisa Bortolotti and Ema Sullivan-Bissett, to whom I am grateful for extensive and very helpful comments on this paper. The paper has also been improved by discussions with my colleagues on the Leverhulme Trust funded project on Bias and Blame, Tom Stafford and Robin Scaife; with Milo Nanni; and from very helpful comments from reviewers for this journal, as well as from extensive and generous comments from Dan Kelly. Special thanks to Federico Picinali for ever fruitful discussions.

References

Adams, R. M. (1985). Involuntary sins. Philosophical Review, 94(1), 3–31.
Anderson, E. (2010). The imperative of integration. Princeton University Press.
Amodio, D. M., & Mendoza, S. A. (2010). Implicit intergroup bias: Cognitive, affective, and motivational underpinnings. In B. Gawronski & B. K. Payne (Eds.), Handbook of implicit social cognition (pp. 353–374). New York: Guilford.
Ashworth, A., & Horder, J. (Eds.). (2013). Principles of criminal law (7th ed.. Oxford University Press.
Bar-Anan, Y., & Nosek, B. (in preparation). A comparative investigation of seven indirect measures of social cognition. Behaviour Research Methods.
Bargh, J. (2005). Bypassing the will: Toward demystifying the unconscious control of social behavior. In Ran R. Hassin, James S. Uleman, & John A. Bargh (Eds.), The new unconscious Oxford series in social cognition and social neuroscience (pp. 37–58). Oxford University Press.
Barnes, A. (1997). Seeing through self-deception. New York: Cambridge University Press.
Brown, J. D. (1986). Evaluations of the self and others: Self-enhancement biases in social judgments. Social Cognition, 4, 353–376.
Calhoun, C. (1989). Responsibility and reproach. Ethics, 99(2), 389–406.
Dehoeve, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. Psychological Bulletin, 135(3), 347–368.
Doris, J. M. (2002). Lack of character: Personality and moral behavior. Cambridge University Press.
Dovidio, J. F., & Gaertner, S. L. (2000). Aversive racism and selective decisions: 1989 and 1999. Psychological Science, 11, 319–323.
Fitzgerald, C. (2013). A neglected aspect of conscience: Awareness of implicit attitudes. Bioethics, 28(1), 24–32.
Fried, J. (2016). The relativism of distance. In Aristotle’s relativism of distance. Aristotle’s relativism of distance. Aristotle’s relativism of distance. Aristotle’s relativism of distance. Aristotle’s relativism of distance. Aristotle’s relativism of distance. Aristotle’s relativism of distance. Aristotle’s relativism of distance. Aristotle’s relativism of distance. Aristotle’s relativism of distance. Aristotle’s relativism of distance.
Garca, J. L. A. (1996). The heart of racism. Journal of Social Philosophy, 27(1), 5–46.
Gawronski, B., Hofmann, W., & Wilbur, C. J. (2006). Are “implicit” attitudes unconscious? Consciousness & Cognition, 15(3), 485–499.
Glaser, J., & Knowles, E. D. (2008). Implicit motivation to control prejudice. Journal of Experimental Social Psychology, 44, 164–172.
Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. Psychological Review, 102, 4–27.
Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2013). Awareness of implicit attitudes. Journal of Experimental Psychology, 10, 274–306.
Holroyd, J. (2012). Responsibility for implicit bias. Journal of Social Philosophy, 44, 274–280.
Holroyd, J., & Kelly, D. (in press). Implicit bias, character & control. In J. Webber, & A. Masala (Eds.), From personality to virtue: Essays in the psychology and ethics of character. Oxford University Press.
Holroyd, J., & Sweetman J. (in press). The heterogeneity of implicit bias. In Saul & Brownstein (Ed.), Implicit bias and philosophy OUP.
Isaacs, T. (1997). Moral context and moral responsibility. Ethics, 107(4), 670–684.
Jost, J. T., Rudman, L. A., Blair, I. V., Carney, D. R., Dasgupta, N., Glaser, J., et al (2009). The existence of implicit bias is beyond reasonable doubt: A reification of ideological and methodological objections and executive summary of ten studies that no manager should ignore. Research in Organizational Behavior, 29, 39–69.
Kelly, D., & Roedder, E. (2008). Racial cognition and the ethics of implicit bias. Philosophy Compass, 3(3), 522–540.
King, M., & Carruthers, P. (2012). Moral responsibility and consciousness. Journal of Moral Philosophy, 9, 200–228.
Levy, N. (2014). Consciousness, implicit attitudes and moral responsibility. Nods, 48(1), 21–40.
Linott, S. (in preparation). Friendship, Homophily, and implicit bias: Friends, for better or worse.
Mele, A. R. (2001). Self-deception unmasked. Princeton University Press.
Monteith, M. J., Voils, C. I., & Ashburn-Nardo, L. (2001). Taking a look underground: detecting, interpreting and reacting to implicit racial biases. Social Cognition, 19(4), 393–417.
Nosek, B. A. (2005). Moderators of the relationship between implicit and explicit evaluation. Journal of Experimental Psychology, 134, 565–584.
Nosek, B. A. (2007). Implicit–explicit relationships. Current Directions in Psychological Science, 16, 65–69.
Fazio, R. H., & Brindl, P. (Eds.). (2009). Attitudes: Insights from the new implicit measures. New York, NY: Psychology Press.
Pronin, Emily, & Kugler, Matthew B. (2007). Valuing thoughts, ignoring behavior: The introspection illusion as a source of the bias blind spot. Journal of Experimental Social Psychology, 43(4), 565–578.
Ranganath, K., Smith, C., & Nosek, B. (2008). Distinguishing automatic and controlled components of attitudes from direct and indirect measurement methods. Journal of Experimental Social Psychology, 44, 386–396.
Saul, J. (2013). Implicit bias, stereotype threat, and women in philosophy. In K. Hutchinson & F. Jenkins (Eds.), Women in philosophy: What needs to change? (pp 39–60). Oxford University Press.

Sher, G. (2009). Who knew?: Responsibility without awareness. Oxford University Press.

Smith, A. M. (2005). Responsibility for attitudes: Activity and passivity in mental life. Ethics, 115(2), 236–271.

Snow, N. (2006). Habitual virtuous action and automaticity. Ethical Theory and Moral Practice, 9, 545–561.

Sullivan-Bissett E. (2015). Implicit bias, confabulation, and epistemic innocence. Consciousness & Cognition, 33, 548–560.

Suls, J., Lemos, K., & Stewart, H. L. (2002). Self-esteem, construal, and comparisons with the self, friends and peers. Journal of Personality and Social Psychology, 82(2), 252–261.

Tidswell, K., Sheeran, P., & Webb, T. L. (in preparation). Self-regulation of the impact of implicit attitudes on behavior (University of Sheffield, 2012).

Vargas, M. (in preparation). Social self-governance, responsibility and implicit bias.

Wallace, R. J. (1994). Responsibility and the moral sentiments. Harvard University Press.

Washington, N., & Kelly, D. (in press). Who’s responsible for this? Moral responsibility, externalism and knowledge about implicit bias. In M. Brownstein, & J. Saul (Eds.), Implicit bias and philosophy. Oxford University Press.