Received: 28 September 2016 | Revised: 30 November 2016 | Accepted: 16 December 2016
DOI 10.1111/phc3.12410

WILEY

ARTICLE

Responsibility for implicit bias[†]

Jules Holroyd | Robin Scaife | Tom Stafford

The University of Sheffield

Correspondence
Dr. Jules Holroyd, Department of Philosophy, The University of Sheffield, 45 Victoria Street, S3 7QB, Sheffield, UK.
Email: j.d.holroyd@sheffield.ac.uk

Funding information
Leverhulme Trust

Abstract

Research programs in empirical psychology from the past two decades have revealed implicit biases. Although implicit processes are pervasive, unavoidable, and often useful aspects of our cognitions, they may also lead us into error. The most problematic forms of implicit cognition are those which target social groups, encoding stereotypes or reflecting prejudicial evaluative hierarchies. Despite intentions to the contrary, implicit biases can influence our behaviours and judgements, contributing to patterns of discriminatory behaviour. These patterns of discrimination are obviously wrong and unjust. But in remedying such wrongs, one question to be addressed concerns responsibility for implicit bias. Unlike some paradigmatic forms of wrongdoing, such discrimination is often unintentional, unendorsed, and perpetrated without awareness; and the harms are particularly damaging because they are cumulative and collectively perpetrated. So, what are we to make of questions of responsibility? In this article, we outline some of the main lines of recent philosophical thought, which address questions of responsibility for implicit bias. We focus on (a) the kind of responsibility at issue; (b) revisionist versus nonrevisionist conceptions of responsibility as applied to implicit bias; and (c) individual, institutional, and collective responsibility for implicit bias.

1 | INTRODUCTION

Research programs in empirical psychology from the past two decades have revealed that implicit biases are pervasive. Implicit biases are typically characterised as automatic associations, of which we may not be aware, that are difficult to control and may conflict with our professed beliefs and values.[1] Although implicit processes are pervasive, unavoidable, and often useful aspects of our cognitions, they may also lead us into error. The most problematic forms of implicit cognition are those which target social groups, encoding stereotypes or reflecting prejudicial evaluative hierarchies. Despite intentions to the contrary, these biases can influence our behaviours and judgements, contributing to patterns of discriminatory behaviour. Here is a paradigmatic example in which implicit biases influence behaviour and judgement:

[†] This paper was produced as part of the Leverhulme Trust Project Grant, on Bias and Blame (RPG-2013-326). We are grateful for their support.

> *Professor P is engaged in anti-racist and feminist activities. She explicitly avows egalitarian values and strives for fair treatment. She believes her students deserve to be treated with equal consideration, and hasn't reflected much on her behaviour, since it seems to her there is no reason for concern. Accordingly, she believes that she adheres to her egalitarian goals, which is what she wants and endorses.*
>
> *However, Professor P harbours implicit biases of the sort found by empirical psychology: associations between whites and intellect (Amodio & Devine, 2006), and negative evaluative associations with black people (Olson & Fazio, 2006). These implicit cognitions influence her behaviour towards her students – again, in ways we might predict given the findings of empirical psychology: her evaluations of the equally good work of black students is slightly less glowing than that of white students (Wood, Hales, Purdon, Sejersen, & Hayllar, 2009); her interactions with black students are beset by micro-behaviours that indicate greater tension and discomfort, and as a result she is marginally less warm and less patient with them than with her white students (Dovidio, Gaertner, Kawakami, & Hodson, 2002).*
>
> *Professor P is not unique – but wholly typical; her colleagues show a similar pattern of evaluations and behavioural interactions, which subtly discriminate against black students. The cumulative effects of this are manifest: fewer black students occupy the top percentiles of classes in her department, fewer pursue or are encouraged to pursue further study in her subject, the student body in her discipline, at the levels at which it is selected for, are disproportionately white.*

In this scenario, implicit biases are complicit in racially discriminatory behaviour, both in individual interactions, and as part of a pattern of behaviours that contribute, alongside other exclusionary factors to sustaining the under-representation of black students in higher education. Note that higher education is not unique; similar scenarios may be found in many domains of education and employment. Implicit racial biases have been found in studies that examine biases in healthcare contexts (Green et al., 2007), housing (Kang, 2014; Turner & Ross, 2007), the criminal justice system (Eberhardt, Goff, Purdie, & Davies, 2004), financial decisions (Tetlock, Kristel, Elson, Green, & Lerner, 2000, see Madva, 2016a for instructive discussion), and in everyday interpersonal interactions (Dovidio et al., 2002). Moreover, other implicit biases target gender, sexuality, age, or disability (Cooley, Payne, Loersch, & Lei, 2014; Levy & Banaji, 2002; Valian, 1999; Wilson & Scior, 2014). It is very likely that any one individual harbours and is influenced by some sort of implicit bias; the findings mean that we cannot rule out that, for each of us, we are influenced in our judgements or behaviours in discriminatory ways.[2]

Such patterns of discrimination are obviously wrong and unjust. But in remedying such wrongs, one question to be addressed concerns responsibility for implicit bias. Yet, unlike some paradigmatic forms of wrongdoing, discriminatory behaviour resulting from implicit bias is often unintentional, unendorsed, and perpetrated without awareness; and the harms may be particularly damaging because they are cumulative and collectively perpetrated. Similarly, the cognitions involved in the biases are ones from which agents may be alienated, because they conflict with endorsed values; the agent may not be aware of their presence in her cognition; nor aware of their role in behaviour. So, what are we to make of questions of responsibility, for the cognitions themselves or for actions influenced by them? In this article, we outline some of the main lines of recent philosophical thought which address questions of responsibility for implicit bias. Our main focus, in keeping with much of the literature, is on responsibility for actions influenced by such biases. These debates have import both for how we think about and respond to implicit bias and for our understandings of responsibility. We focus on (a) the kind of responsibility at issue; (b) revisionist versus nonrevisionist conceptions of responsibility as applied to implicit bias; and (c) individual, institutional, and collective responsibility for implicit bias. We close with a practical example of institutional responsibility and implicit bias training that animates some of these issues.

2 | WHAT SENSE OF RESPONSIBILITY?

In remedying wrongs, we might be interested both in causal responsibility and moral responsibility; that is, we want to know both who or what brought about the wrong and who is to blame for it. Notably, some have argued that attention to individuals and their implicit biases is inapt in addressing these patterns of discrimination, because the causal responsibility for such inequalities and discriminatory patterns is fundamentally structural (Anderson, 2012; Haslanger, 2015). Accordingly, remedies would do better to target these fundamental, structural causes.

On the other hand, others have maintained that implicit biases, and crucially, actions that manifest them, *are* causally responsible for some of these harms and inequalities (Saul, 2013a)—and, attention should be paid to understanding the role of implicit bias in sustaining these harms and attempting to mitigate their influence.

However, in drawing attention to the role of implicit biases in sustaining harms and exclusion, it has been contentious whether individuals are always morally responsible for these harms. Saul (2013a) has suggested not, writing that we should

> abandon the view that all biases against stigmatised groups are *blameworthy* … [because a] person should not be blamed for an implicit bias that they are completely unaware of, which results solely from the fact that they live in a sexist culture. (2013a: 55, emphasis in original)

Here, Saul (2013a) focuses on whether individuals are responsible merely for having implicit biases. However, this question is distinct from the question of whether one is responsible for manifesting implicit bias in one's actions, and one may be blameworthy for one's actions even if one is not blameworthy for the presence of the implicit association in one's cognitive make up (Holroyd, 2012). In any case, Saul (2013a) suggests that once individuals are aware that they have, and may be influenced by, implicit biases, they are responsible for doing something about it:

> They may, however, be blamed if they fail to act properly on the knowledge that they are likely to be biased—e.g., by investigating and implementing remedies to deal with their biases (ibid.)

Saul's (2013a) remarks here exemplify both the backward-looking concern with appraising individuals for their cognitions and behaviours; and attention to the forward-looking issues concerned with taking responsibility for correcting harms and mitigating costs (of course, failure to meet these forward-looking obligations may render one blameworthy, in the backward-looking sense, for that failure). In the wider debate, theorists have teased apart the forward- and backward-looking concerns into (*inter alia*) the following different questions, each of which might legitimately be regarded as concerning a sense in which the agent is responsible:

a. Does the attitude or act *reflect badly* (or well) on the agent; is there are fault (or indeed credit) that can be attributed to her?
b. Should the agent be regarded as *blameworthy* for the fault she has, or has demonstrated in her action; should she bear some cost or burden (in the form of sanction or blame) for this?

These two questions are backward-looking and contrast with a third concern, namely,

c. What forward-looking obligations do individuals have for dealing with the fault, or problematic behaviour?

Those familiar with the responsibility literature will be aware that the two backward-looking concerns, above, have sometimes been characterised as questions of attributability (a) and accountability (b) (cf. Smith, 2012; Watson, 1996). However, these two terms have (confusingly) garnered somewhat different usage in the literature on responsibility for implicit bias. Zheng (2016) has labelled the backward-looking senses (a and b) as together questions of attributability and characterised as accountability the forward-looking issue of how remedial obligations are distributed (this way of using the two notions, Zheng, 2016 notes, has more in common with the way Scanlon uses the terms).

All three questions are activated in thinking about responsibility for implicit bias. To avoid terminological confusion, in the following discussion, we make reference to the particular question at the issue: does it reflect on the agent; is she blameworthy for it (the backward-looking questions); do obligations to address bias attach to an agent (the forward-looking concerns)?

Zheng (2016) argues that we arrive at different senses of responsibility via different routes. The backward-looking questions concern what reflects well or badly on the agent, such that they are subject to “appraisive responses” such as praise or blame (or perhaps other evaluations of character). Blameworthiness, thus construed, is a backward-looking matter. For example, to find Professor P responsible in this sense would be to suppose that her discriminatory evaluations and behaviours flow from the relevant features of her agency—some metaphysical or psychological facts—which allow those actions to reflect on her as an agent and enable us to perhaps find her blameworthy for those harmful acts she perpetrates.

In contrast, the forward-looking concerns (what Zheng, 2016 calls accountability) track not deep facts about the agent and what reflects on her, but rather social facts about the distribution of benefits and burdens when something has gone wrong: who should step in to deal with the costs of wrongdoing, irrespective of whether they were at fault. We might hold Professor P to be charged with acting to remedy the costs of the harms she is implicated in, irrespective of whether we think those harms are ones that reflect on her, or for which she is blameworthy.

In some cases of implicit bias, it may be clear that the biases warrant backward-looking and appraisive responses, for example, in cases where the agent has cultivated the biases, this would reflect badly on her, and she may be blameworthy for her biases. Likewise, in cases in which the agent endorses or would endorse the discriminatory behaviour (Holroyd, 2016; Zheng, 2016), the behaviour appears to reflect a fault on her part; moreover, judgements of blame and expressions of blame—may also be engaged.

However, in other cases—which meet the paradigm illustrated by Professor P—things may not be so clear. Various authors have suggested that there are reasons for which the discriminatory behaviour may not reflect badly on the agent, and certainly not in a way that warrants blame (see, e.g., Saul, 2013a, quoted above; Levy, 2012, 2014 and Section 3 below), because the behaviour is not something the agent is aware of, nor under the agent's control. On the other hand, Brownstein sets out the case for taking actions that express implicit bias to be reflective of who the agent is, insofar as they reflect the agent's cares (which may not form a coherent set, and hence may conflict with her other cares). What the agent cares about, Brownstein suggests, is reflected in multitrack patterns of behaviour that reveal the agent's dispositions (2015: 12). As such actions that express implicit bias can be attributable to the agent and may ground moral appraisals of her.

In any case, one might hold such agents responsible in the in the sense of having forward-looking obligations to deal with the problem, irrespective of whether bias, or actions influenced by it, reflects badly on them or is something for which they can be blamed (Zheng, 2016). Withholding backward-looking appraisive judgements and maintaining that the agent is responsible only in the forward-looking sense of having remedial obligations may seem appealing for three reasons. First, one may have doubts about appraising an agent yet insist that it is important that someone is held responsible nonetheless (though see Holroyd, 2012, Brownstein, 2015, and the discussion in Section 3 below, for arguments addressing different versions of such doubts). As Zheng (2016: 74) puts it, it is important that we are held to forward-looking responsibilities “because it is appropriate for us to clean up after our own actions when a mess has been made,” even if the mess does not reflect badly on us, or warrant blame. When someone has been harmed by implicit bias, she is owed “compensation, apology, and redress” irrespective of whether anyone is found to be at fault or blameworthy (2016: 74). (Note that it remains for proponents of this strategy to delineate senses of apology and redress that can play this function, whilst being distinct from those appraisive responses characteristic of backward-looking responsibility judgements.) Second, it may be difficult to identify, with sufficient certainty to warrant blame, whether implicit bias has played a role in discriminatory behaviour in any isolated case.

A third reason for focusing on responsibility in the forward-looking sense concerns worries about the engagement of the appraisive responses. Saul (2013a) worries that blame might produce backlash or hostility. This is a common concern: Zheng (2016:80) marshals evidence (Legault, Gutsell, & Inzlicht, 2011), which suggests that high threat confrontations can produce defensiveness and greater bias (implicit and explicit); Vargas (2017) raises worries about securing the needed buy-in for norms against bias, if agents experience backlash in response to blame that is not accepted or well received.

The appeal of a focus on holding people responsible in the exclusively forward-looking sense, for actions influenced by implicit bias, rests on (i) the extent to which we are convinced that agents are not (or cannot be identified as) responsible in the backward-looking senses (more on this in Section 3 below); and (ii) the plausibility of concerns about negative responses to appraisive evaluations such as blame.

On this latter point, it is worth noting that appraisive responses may include a wide range of moral evaluations that need not be characterised by “high threat” features (such as high emotion, or challenge to the agent's self-conception or esteem). For a start, appraisals that focus on whether biases reflect badly on the agent (sense (a) above) may not involve blame at all. Rather, they might invoke an evaluative judgement about the agent and her character—she is cruel, or she is racist—without taking a stance on whether this is her fault. This is the sort of “aretic appraisal” (Watson, 1996) that has been taken to be distinct from judgements of blameworthiness or expressions of blame. But even blaming responses may not be high threat. Instead, preliminary empirical work by Malle et al. (2014) suggests that individuals are more likely to see responses as forms of blaming when they are thoughtful, calm, and delivered without high emotion (in contrast, see the conception of blame outlined and critiqued by Pickard, 2013, which supposes blame has an emotional “sting”). Finally, some initial work on reactions to blame for implicit racial bias has found that blaming responses did not increase implicit biases and in fact significantly increased individuals' *explicit* intentions to take action to address bias. This effect on explicit intentions was long lasting, persisting in a 6 month retest, in which participants continued to report stronger motivations to tackle discrimination (see Scaife, Holroyd, Stafford, & Bunge, (ms.)).[3] In short, no “backlash effect” was observed. This result may seem surprising at first. But note that it is less so if worries about blame for implicit bias rest on a mistaken paradigm of blaming communications: one that supposes that individuals are being told they are bad and wrong. Instead, blame for implicit bias might be better modelled—at least in some cases—as a communication in which individuals are alerted to the fact that they have violated a norm that they already (rightly) endorse (see Scaife et al., ms. for more discussion). We might expect reactions to these different sorts of communication to elicit rather different responses. It seems that considerable work is still to be done on understanding how to model different sorts of moral response and on empirically evaluating how they are received and reacted to.

If responses such as blame are effective in influencing individuals' attitudes towards avoiding implicit bias, then there may be reason to deploy such responses irrespective of whether individuals in fact deserve such responses. But absent desert, the justificatory burden for blame is certainly greater. So, much hinges on whether agents are in fact blameworthy for implicit bias or implicitly biased actions. This depends on how we conceive of the conditions for this kind of responsibility, to which we now turn.

3 | FOLK VERSUS REVISIONIST CONCEPTIONS OF RESPONSIBILITY

Many philosophical conceptions of responsibility aim to articulate, perhaps with some refinements, the conditions of responsibility embedded within and deployed by folk conceptions of responsibility. Whilst there is of course much long standing philosophical disagreement over how to articulate and unpack this conception, many agree on the centrality of the following notions in determining moral responsibility, in the sense of blameworthiness, for attitudes and actions based upon them: awareness or knowledge conditions, and control conditions (which authors have spelled out in terms of reasons-responsiveness or evidence-sensitivity conditions that have featured heavily in the literature

[Fischer & Ravizza, 2000; Levy, 2014, forthcoming]). In a paradigm case of wrong-doing for which the agent is morally responsible, the agent's attitudes are under her control in that they are suitably responsive to reasons and to evidence; she is aware of what she is doing and of the consequences of her actions; and she is able to ensure her actions reflect her endorsed values.

It is clear enough that the discriminatory actions influenced by implicit bias—such as the evaluations and interactions of Professor P—do not meet this paradigm. Professor P′s implicit biases are not sensitive to her other (evidentially supported) attitudes, beliefs and values, but conflict with them; she is unaware that she discriminates; and she does so automatically, without the reflective or deliberative control we are sometimes able to exercise. This departure from the paradigm cases of moral responsibility has lead some authors to claim that individuals are not responsible for actions influenced by implicit bias: such actions cannot reflect badly on them, nor are they always blameworthy for them (see Saul, 2013a above).

Levy offers extensive arguments for the claim that individuals such as Professor P are not blameworthy for implicitly biased actions, because her actions are not actions of which she is appropriately consciously aware;[4] nor is she able to exercise reflective control over them; nor are the attitudes integrated with (or responsive to) her endorsed evaluative attitudes (Levy, 2012, Levy, 2014, forthcoming). Accordingly, Levy claims, they cannot be reflective of who the agent is, and so it cannot be appropriate to respond with certain burdens (sanction, blame, and punishment; forthcoming: 4).

But of course, it is debatable whether the notions of awareness, or control, deployed in such arguments are the relevant ones for blameworthiness (see Holroyd, 2012), and some lines of debate have engaged these issues.

Awareness

Consider first the awareness requirement. The idea that we are responsible (in the sense of blameworthy for) only for those attitudes or actions that fall within, or are guided by, conscious awareness faces sustained critical challenge from Sher (2009). Such a condition would implausibly render us exculpated from (inter alia) any instance of absent-mindedness, inattention, or forgetting (cf. Smith, 2005). We should instead, Sher suggests, start from the thought, that "when someone acts wrongly or foolishly, the question on which his responsibility depends is not whether he *is* aware that his act is wrong or foolish, but rather whether he *should* be" (Sher, 2009:20).

In the context of implicit bias, various senses of awareness are in circulation (taxonomised Holroyd, 2014): agents may have (or lack) introspective awareness of their implicit biases; individuals may have (or lack) inferential awareness of their proneness to implicit biases, that is, an awareness based on inferences drawn from empirical studies to our own propensities. Third, individuals may have (or lack) observational awareness, that is, awareness based on the observation of their own behaviours, in one-off cases or in patterns of action.[5] Relatedly, individuals may be aware (or not) of the moral significance of their action (Levy, 2014). Combined with Sher's insight, Holroyd argues that responsibility will depend not on whether individuals in fact have or lack any such sense of awareness, but rather on whether we should have any of these kinds of awareness; and in the instance in which we do not, whether any such failure is culpable.

For instance, such a strategy is deployed by Washington and Kelly (2016), who argue that agents in certain positions of power or authority—those on hiring panels, say—*should have* awareness of their own likely implicit biases. That is, given their role responsibilities and the impact of their decisions, they should be aware of the relevant research on implicit bias, and be able to make inferences about the ways in which their own judgements may be biased—and, crucially, the steps they may take to mitigate this. The issue is then whether such individuals are culpable when they lack such awareness.

Washington and Kelly (2016) take this to support "externalist" conditions on responsibility, such that an individual's failing to be aware of what she should be may or may not be culpable, contingent upon her epistemic environment. In a context, in which information—from which inferences can be drawn—is not readily available, the failure of awareness is not (or at least, is less) culpable. But in an epistemic environment such as ours, in which such information is—increasingly—more readily available, we are culpable for lacking awareness of the information of which we should, and can reasonably be expected to be, aware.

Holroyd (2014) argues that in fact, we need not demand even this inferential awareness; agents may have observational awareness even in epistemic environments in which research on implicit bias is not "common knowledge." This can be gained by careful reflection on one's own behaviour and its role in sustaining patterns of systemic injustice. Although the relevant observations may be more easily made when information about implicit bias is prevalent, such information is not necessary. Indeed, prior to the advent of research on implicit bias, plenty of testimonial evidence from individuals targeted by implicit racial biases could have prompted such reflections and observations (see Holroyd & Puddifoot, forthcoming). And, insofar as these observations are blocked by motivated ignorance—avoidance or ignoring (perhaps subconsciously) of evidence that serves one's goals—or self-deception, or excessive weight given to misleading introspective evidence, a failure to have this sort of awareness is culpable to some degree.

In sum, on the folk conception of responsibility, some awareness conditions may be required for individuals to be blameworthy for their wrongful actions. But it is far from clear that simply lacking conscious awareness of facts about implicit bias, or the fact that one is discriminating, is exculpatory. Rather, we might expect that individuals—at least in some epistemic contexts—have, and should have, certain kinds of awareness about their propensity to harbour and manifest implicit bias.

Control

Contention has also arisen over the kind of control necessary to ground blameworthiness for actions influenced by implicit biases. As Saul (2013a, 2013b) has argued, individuals lack direct control over the activation and influence of implicit bias. We cannot simply will that biases do not affect us, and make it so. Indeed, studies have indicated that trying to suppress implicit bias has a rebound effect (Follenfant & Ric, 2010).

However, many aspects of our cognition and action for which we are held responsible are not under our direct control. This suggests that other forms of control are relevant to moral responsibility. Two questions then arise: what forms of control are necessary for moral responsibility? And do we have these forms of control with respect to implicitly biased actions?

Levy (2014, forthcoming) has argued that agent's attitudes must be under a certain kind of rational control—they must be inferentially sensitive to, and integrate with, the other evidence-sensitive attitudes of the agent. Drawing on evidence that implicit attitudes are not inferentially sensitive in the appropriate way, he argues that actions influenced by implicit biases cannot be reflective of the agent in a way that renders her liable to blame (Levy, 2014 pp.812–816). Such implicit attitudes cannot, Levy (2014, forthcoming) argues, be properly integrated into the agent.

It is worth noting, however, that any such lack of inferential sensitivity does not necessitate a lack of integration. Implicit biases may accord with and thereby reflect the agent's values, even if not under rational control (Holroyd, 2016). The paradigm case of this would be an explicit racist who nonetheless has implicit cognitions and automatic responses, which contribute to (and function in the service of) his morally repugnant goals.

In any case, rational control may not be necessary for moral responsibility: other sorts of control may suffice in order for agents to meet any control conditions for moral evaluation. For example, Snow (2005) identifies a sort of 'intervention control' that agents may exercise over automatic processes (in the context of discussing automaticity, rather than implicit biases per se—see Holroyd and Kelly, 2016 for discussion of intervention control in the context of implicit biases). Although the processes themselves are not rationally guided but are automatized and habituated, agents may intervene to halt or redirect these processes. Such control may be possible in the case of implicit biases. For example when on a hiring panel, Professor P may intervene to halt the automatic influence of implicit biases, by, say, referring to a checklist of previously agreed upon criteria, as a strategy to try to overcome any tendency to bias (cf. WISELI, 2012, p.6).

Other forms of indirect control may also be deployed, such as retraining or conditioning one's own cognitions via techniques such as "implementation intentions." This technique involves conditioning certain counter-stereotypical automatic thoughts in response to environmental cues, for example, automatizing the thought "safe," as prompted by any black person, in order to try to overcome associations between blackness and danger or hostility (Stewart and Payne, 2008). Agents may add to these cognitive props environmental props also—reshaping their social environment in a way that impacts on their cognitions in desirable ways. For example, some studies have suggested that the presence of pictures depicting counter-stereotypical exemplars reduces negative biases (Blair, 2002).

These modes of indirect control—manipulating one's environment or one's cognitions in order to secure desirable patterns of thought and behaviour—have been identified by Holroyd and Kelly (2016) as forms of "ecological control." They argue that this kind of control is in fact mundane, oft-deployed (e.g., consider organising one's office to ward off procrastinatory tendencies), and sufficient for meeting control conditions for moral responsibility.

Although the kinds of ecological control so far outlined involve intentionally deploying strategies to exert control, Holroyd and Kelly (2016) argue that this is not necessary. One form of ecological control outlined is that of unintentional indirect control: an example of this is the sort of control demonstrated by participants in Moskowitz and Li's (2011) studies, whereby agents with strong egalitarian commitments appeared to be better able to block the activation of implicit biases. This was not explicitly intended by those agents, but seemed to be an automatic function of their strongly held egalitarian commitments. Because agents who lacked such commitments were less effective at blocking the activation of implicit stereotypes, Holroyd and Kelly (2016) argue, it is to the agent's credit and attributable to her, when her values better enable the pursuit of her goals.

Whether agents who are influenced by implicit biases meet control conditions for moral responsibility, then, will depend on what notion of control one teases out of the folk conception, how defensible this notion of control is, and the extent to which it is exercised with respect to implicit biases. Note that in teasing out these notions, there is some refinement of the ideas found in the folk conception: in the articulation of the idea of "ecological control," for example. But authors proposing these notions take such ideas to be latent in, or extensions of, the folk concept of moral responsibility. This contrasts with the strategies deployed by theorists who are explicitly revisionist about moral responsibility.

Revisionism

Some authors have suggested that thinking about blameworthiness for implicit biases should motivate revisions to this understanding of the concept of moral responsibility, whereby it is acknowledged explicitly that such revisions involve a departure from seemingly intuitive thought about what it is for attitudes and actions to reflect on us, and render us liable to blame.

Two authors who propose such revisions are Glasgow (2016) and Faucher (2016). Each is motivated by the idea that we are morally responsible for actions influenced by implicit biases—but that this can only be adequately captured by revising our analyses of moral responsibility. For example, Glasgow suggests that whilst common sense suggests that the conditions for moral responsibility are invariant—they remain constant across agents and contexts—his reflections on implicit bias lead to the revisionary conclusion that they are variant. In particular, the conditions vary depending on the content of the attitude held and the action performed on the basis of it (2016: 48; later, Glasgow suggests that the content is a proxy for the kind or degree of harm caused [2016: 56]). On this view, the content of the attitude determines whether an agent's alienation from her attitude (or action performed on that basis) suffices to exculpate. In the case of Frankfurt's alienated drug user, alienation seems to suffice. But in cases where the attitude involves a relational harm—the attitudes that drive infidelity or prejudicial attitudes involved in implicit biases—the fact that the agent is alienated from her attitude does not serve to exculpate.

Faucher (2016) also argues for a revisionary form of variantism about moral responsibility. On his view, the dimension of variation concerns the conditions that are applied depending on one's position in relation to the harm. If one is—or has been—a victim of discrimination, the conditions one deploys in evaluating moral responsibility will not make reference to explicit or conscious intentions, he argues. Harm has been perpetrated and it matters not, from this point of view, whether it was consciously intended. However, from the point of view of perpetrators, conscious intention does matter.

One might think that this difference rests on mistake or disingenuousness on the part of the perpetrators. But if it is genuine and withstands reflective scrutiny, then our conditions for moral responsibility may turn out to be variant—and an invariantist articulation of the concept of moral responsibility should be revised accordingly.

Faucher (2016) also suggests that other revisions; first, to the notion of control necessary for responsibility. As discussed above, forms of indirect control (Faucher, 2016 calls those which involve retraining of one's cognitions a kind of “bottom-up” control) may be sufficient for moral responsibility. And notions of responsibility that make reference to the “real self”—the part of the self that reveals where the agent stands—may have to account for the agent's real self as including some unendorsed implicit attitudes, as well as her endorsed explicit attitudes.

An explicitly revisionist model of responsibility has been developed and defended by Manuel Vargas in recent years (2005, 2009, and 2013). His recent work has applied this model to implicit biases. Vargas (2017) sees responsibility judgments as part of a practice that can be justified if it serves certain forward-looking goals of cultivating moral agency. On this view, to be a responsible agent is to stand in certain social relations (rather than to meet certain metaphysical conditions). Thus construed, it is crucial that the agent's context supports moral agency in relation to a certain set of considerations. Vargas worries that our current environment (the “moral ecology”) does not yet provide the right sort of support. We are perhaps not yet suitably cognizant of, or sensitive to considerations of implicit bias, to make it appropriate—or fruitful—to hold each other responsible or liable to blame, on his view. But, as our moral ecology evolves, it may be appropriate to do so. This line of argument has resonance with Washington & Kelly's (2016) externalist conditions for moral responsibility, whereby what can reasonably be expected of agents may shift as epistemic environments change. Note also that this view about the current impropriety of blame shares much with the concerns about the defensiveness and hostility that such appraisive responses may invoke. Yet, as we have seen, it is not at all clear that we do find these responses—it may instead be that these interpersonal interactions of blame are part of the tools that shape the moral ecology and increase sensitivity to, and motivation to combat, implicit biases (Holroyd 2012; Scaife et al., ms.).

These moral responses are premised chiefly on the idea that *individuals* may be held responsible for implicit biases. But this is not the only option.

4 | INDIVIDUAL, INSTITUTIONAL, AND COLLECTIVE RESPONSIBILITY

So far, the focus has been on holding individuals morally responsible for particular discriminatory actions or judgements made under the influence of implicit bias. But as noted at the outset, the problem is not simply one of particular discriminatory behaviours but also of their role in sustaining patterns of marginalisation, exclusion, and hierarchy. Insofar as these social structures are collectively caused, sustained, and stand in need of collective remedy, it appears that an approach that requires collective responsibility is also needed.

Of course, questions remain about who is collectively responsible for what, in both the backward- (evaluation of character or identification of who is to blame) and forward-looking (who bears remedial costs) senses. One might maintain that all of us are to blame for, or have a role in remedying, the harms that result from implicit biases. Indeed, as Jacobsen has noted (2016, 174), agents need not have done anything particularly wrong in order to have been complicit in sustaining patterns of exclusion. This complicity may justify being held accountable for remedying the harms. Or one might maintain that responsibility and blameworthiness falls collectively, but proportionally, on those groups who have greater role in perpetrating harm (cf. Washington and Kelly's (2016) remarks about people in “gatekeeper” positions having greater responsibilities). Likewise, in terms of forward-looking responsibility, Sie and Van Voorst Vader-Bours suggest that “it seems reasonable to expect those in positions of power to take more responsibility for change” (2016: 107). Moreover, if it is right that remedying implicit biases is most effectively done by making broader structural changes (Anderson, 2012; Haslanger, 2015; Saul, 2013b), then certainly institutional and collective action will be needed to enact this. But as Madva (2016b) has persuasively argued, attention to individual or collective endeavours should not be seen as competing; rather, the complex interplay between individual and more collective or structural interventions needs to be recognised. For example, institutional change requires individual buy-in and motivation to instigate change. We close with an example of how both institutional and individual responsibility needs to have a role in practical efforts to remedy implicit bias.

Many institutions now offer "implicit bias training," seminars that aim to make individuals within an institution aware of the possibility of implicit biases influencing their behaviour, and of remedies that could be undertaken to combat this. One model for thinking about these training sessions is in terms of individual responsibility: individuals are made aware of certain problematic facts (their propensity for bias) and are given certain tools that they can apply to their cognitions or to their workplace procedures, in order to try to prevent their biases from having a role. This training devolves (forward-looking) responsibility to individuals for remedying bias.

However, this model may be flawed, both in terms of efficacy and theoretical warrant. First, if the emphasis is on individual de-biasing, this requires that individuals accept and acknowledge that they themselves have implicit bias. Because cognitive biases include objectivity bias (the propensity to believe that one is more objective than one's counterparts [Pronin, Gillovich, & Ross, 2004; Rachlinski, Johnson, Wistrich, & Guthrie, 2009]), there are significant obstacles to securing this acceptance. Moreover, we need not (and often cannot) pinpoint exactly whom is biased and on what occasions; it suffices to motivate institutional change to know that many individuals will, on many occasions, display certain kinds of biases. Instead, the emphasis should be on the fact that bias is pervasive and as individuals within institutions we should seek procedures that make those institutions robust against the influence of implicit bias. Second, this model may seem to imply that the primary focus of addressing implicit biases is to de-bias individuals. If this is the aim, then devolving responsibility to individuals is a flawed strategy: because there exist few studies that show the long-term efficacy of changes to implicit biases (Lai et al., 2014, though cf. Madva, ms.). The point is not that individuals should not try to de-bias (see Madva, ms. for the claim that such efforts, are, in the grand scheme of things, rather small), nor that individual attitude change is not relevant at all; but that such efforts should not be independent of institutional change, which itself requires that individuals are motivated to institute and sustain those changes. Moreover, and fundamentally, this model risks obscuring the fact that individual de-biasing is only part of, and perhaps instrumental to, the central goal, which is to address patterns of discriminatory outcome. This may be secured by means other than individual de-biasing.

The overall point, then, is that in recognising the role that individuals and institutions must play, we see that institutional responsibility must reach further than simply providing implicit bias training, on the assumption that this devolves responsibility to individuals for dealing with discrimination. Rather, the responsibility is with (individuals within) institutions to take sustained measures to address whatever mechanisms are producing discriminatory outcomes. Implicit bias training may be a part of those measures, but institutional change must extend well beyond this.

ENDNOTES

[1] Not all parties to the debate accept this characterisation. See, for example, Mandlebaum, 2015 and Levy, 2014 for the claim that implicit biases likely have propositional structure.

[2] For debate over how likely this is, see Jost et al., 2009; Greenwald, Poehlman, Uhlmann, & Banagi, 2009; and Oswald, Mitchel, Blanton, Jaccard, & Tetlock, 2013.

[3] Note that these findings address one objection to communicating blame but the issue of blameworthiness and whether one should communicate blame may be thought to come apart (though for some consequentialists, to be blameworthy just is for there to be sufficient reason to blame). There might be reasons (backlash) to withhold blame even if the agent is blameworthy; we can imagine scenarios in which one should blame (if the gains are great) even if the agent is not blameworthy.

[4] See Cameron, Payne, & Knobe, 2010 for studies that suggest that Levy's claims have support in the folk conception of moral responsibility. Terminologically, it is important to note that Levy is concerned with blameworthiness, even though he frames his question as concerning whether biases are "attributable" to the agent (forthcoming p.4).

[5] Note, moreover, that it is a mistake to suppose a failure to report implicit biases (on self-report measures) is equivalent to lacking awareness in any one of these senses (since there may be various reasons for failures of self-report). See Stafford, 2014 for discussion.

WORKS CITED

Amodio, D., & Devine, P. (2006). Stereotyping and evaluation in implicit race bias: Evidence for independent constructs and unique effects on behavior. *Journal of Personality and Social Psychology*, *91*(4), 652–661.

Anderson, E. (2012). Epistemic justice as a virtue of social institutions. *Social Epistemology*, *26*(2), 163–173.

Blair, I. (2002). The malleability of automatic stereotypes and prejudice. *Personality and Social Psychology Review*, *3*, 242–261.

Brownstein, M. (2015). Attributionism and moral responsibility for implicit bias. *Review of Philosophy and Psychology*, 1–22. doi:10.1007/s13164-015-0287-7

Cameron, C. D., Payne, B. K., & Knobe, J. (2010). Do theories of implicit race bias change moral judgments? *Social Justice Research*, *23*(4), 272–289. doi:10.1007/s11211-010-0118-z

Cooley, E., Payne, B. K., Loersch, C., & Lei, R. (2014). Who owns implicit attitudes? Testing a metacognitive perspective. *Personality and Social Psychology Bulletin*, *41*(1), 103–115.

Dovidio, J. F., Gaertner, S. E., Kawakami, K., & Hodson, G. (2002). Why can't we just get along? Interpersonal biases and interracial distrust. *Cultural Diversity and Ethnic Minority Psychology*, *8*(2), 88.

Eberhardt, J., Goff, P., Purdie, V., & Davies, P. (2004). Seeing black: Race, crime, and visual processing. *Journal of Personality and Social Psychology*, *87*(6), 876–893.

Faucher, L. (2016). Revisionism and moral responsibility for implicit attitudes. In M. Brownstein, & J. Saul (Eds.), *Implicit bias and philosophy* (Vol. 2) (pp. 115–145). Oxford, UK: Oxford University Press.

Fischer, J., & Ravizza, M. (2000). *Responsibility and control: a theory of moral responsibility*. New York, USA: Cambridge University Press.

Follenfant, A., & Ric, F. (2010). Behavioral rebound following stereotype suppression. *European Journal of Social Psychology*, *40*, 774–782.

Glasgow, J. (2016). Alienation and responsibility. In M. Brownstein, & J. Saul (Eds.), *Implicit bias and philosophy* (Vol. 2) (pp. 37–61). Oxford, UK: Oxford University Press.

Green, A., Carney, D., Pallin, D., Ngo, L., Raymond, K., Lezzoni, L., & Banaji, M. (2007). Implicit bias among physicians and its prediction of thrombolysis decisions for black and white patients. *Journal of General Internal Medicine*, *22*, 1231–1238.

Greenwald, A., Poehlman, T. A., Uhlmann, E. L., & Banagi, M. R. (2009). Understanding and using the implicit association test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, *97*, 17–41.

Haslanger, S. (2015). Distinguished lecture: Social structure, narrative and explanation. *Canadian Journal of Philosophy*, *45*(1), 1–15.

Holroyd, J. (2012). Responsibility for implicit bias. In M. Crouch, & L. Schwartzman (Eds.), *Journal of social philosophy*, *special issue: gender, implicit bias and philosophical methodology* (Vol. 43) (pp. 274–306).

Holroyd, J. (2014). Implicit bias, awareness and imperfect cognition. In L. Bortolotti, & E. Sullivan-Bissett (Eds.), *Consciousness and cognition, special issue: costs and benefits of imperfect cognitions* (pp. 511–523).

Holroyd, J. (2016). What do we want from a model of implicit cognition? *Proceedings of the Aristotelian Society*, *116*(2), 153–179.

Holroyd, J., & Kelly, D. (2016). Implicit bias, character and control. In J. Webber, & A. Masala (Eds.), *From personality to virtue: essays on the philosophy of character* (pp. 106–134). New York, USA: Oxford University Press.

Holroyd, J., & Puddifoot, K. (forthcoming). Implicit bias and prejudice. In M. Fricker, P. J. Graham, D. Henderson, N. Pedersen, & J. Wyatt (Eds.), *Routledge handbook of social epistemology*. Abingdon, UK: Routledge.

Jacobsen, A. (2016). Reducing racial bias: attitudinal and institutional change. In M. Brownstein, & J. Saul (Eds.), *Implicit bias and philosophy* (Vol. 2) (pp. 173–190). Oxford, UK: Oxford University Press.

Jost, J., Rudman, L., Blair, I., Carney, D., Dasgupta, N., Glaser, J., & Hardin, C. (2009). The existence of implicit bias is beyond reasonable doubt: A refutation of ideological and methodological objections and executive summary of ten studies that no manager should ignore. *Research in Organisational Behaviour*, *29*, 39–69.

Kang, J (2014) Facing the enemy: Segregation and implicit bias. http://furmancenter.org/research/iri/essay/implicit-bias-and-segregation-facing-the-enemy accessed 19-09-2016

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J. L., Joy-Gaba, J. A., … Nosek, B. A. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General*, *143*, 1765–1785.

Legault, L., Gutsell, J. N., & Inzlicht, M. (2011). Ironic effects of anti-prejudice messages: How motivational interventions can reduce (but also increase) prejudice. *Psychological Science*, *22*(12), 1472–1477.

Levy, N. (2012). Consciousness, implicit attitudes, and moral responsibility. *Noûs*, *48*, 21–40.

Levy, N. (2014). Neither fish nor fowl: Implicit attitudes as patchy endorsements. *Noûs*. doi:10.1111/nous.12074

Levy, N., (forthcoming) "Implicit bias and moral responsibility: Probing the data" *Philosophy and Phenomenological Review*.

Levy, B. R., & Banaji, M. R. (2002). Implicit Ageism. In T. D. Nelson (Ed.), *Ageism: stereotyping and prejudice against older persons*. (pp. 49–75). Cambridge, MA: MIT Press.

Madva, A. (2016a). Virtue, social knowledge, and implicit bias. In Brownstein, & Saul (Eds.), *Implicit Bias and Philosophy* (Vol. 1) (pp. 191–215). Oxford, UK: Oxford University Press.

Madva, A. (2016b) "A plea for anti-anti-individualism: How over-simply psychology misleads social policy *Ergo* ms. available at: http://alexmadva.com/sites/default/files/Madva%20-%20Plea%20for%20Anti-Anti-Individualism%202015.11.16.pdf accessed 19-09-2016

Madva, A. (ms.), "Biased against de-biasing: On the role of (institutionally sponsored) self-transformation in the struggle against prejudice" http://alexmadva.com/sites/default/files/Biased%20Against%20Debiasing%202015.4.pdf accessed 19-09-2016

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, *25*(2), 147–186.

Mandlebaum, E. (2015). Attitude inference and association: On the propositional structure of implicit bias. *Nous*. doi:10.1111/nous.12089

Moskowitz, G. B., & Li, P. (2011). Egalitarian goals trigger stereotype inhibition: A proactive form of stereotype control. *Journal of Experimental Social Psychology*, 172–179.

Olson, M. A., & Fazio, R. H. (2006). Reducing automatically activated racial prejudice through implicit evaluative conditioning. *Personality and Social Psychology Bulletin*, *32*, 421–433.

Oswald, F. L., Mitchel, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology*, *105*(2), 171–192.

Pickard, H. (2013). Irrational blame. *Analysis*, *73*(4), 613–626.

Pronin, E., Gillovich, T., & Ross, L. (2004). Objectivity in the eye of the beholder: Divergent perceptions of bias in self versus others. *Psychological Review*, *111*(3), 781–799.

Rachlinski, J.J., Johnson, S.; Wistrich, A.J., and Guthrie, C. (2009), "Does unconscious racial bias affect trial judges?" (2009). Cornell Law Faculty Publications. Paper 786. http://scholarship.law.cornell.edu/facpub/786

Saul, J. (2013a). Implicit bias, stereotype threat, and women in philosophy. In F. Jenkins, & K. Hutchinson (Eds.), *Women in philosophy: what needs to change?* (pp. 39–60). New York, USA: Oxford University Press.

Saul, J. (2013b). Implicit bias and scepticism. *Disputatio*, *5*(37), 243–263.

Scaife, R. Holroyd, J., Stafford, T., Bunge, A., (ms.) The effects of moral interactions on implicit racial bias. https://osf.io/eubjp/

Sher, G. (2009). *Who knew? Responsibility without Awareness*. New York, USA: Oxford University Press.

Sie, M., & Van Voorst Vader-Bours, N. (2016). Stereotypes and prejudices: whose responsibility? Indirect personal responsibility for implicit biases. In M. Brownstein, & J. Saul (Eds.), *Implicit Bias and Philosophy* (Vol. 2) (pp. 90–114). Oxford, UK: Oxford University Press.

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

Smith, A. (2012). Attributability, answerability and accountability: In defence of a unified account. *Ethics*, *122*, 575–589.

Snow, N. (2005). Habitual virtuous actions and automaticity. *Ethical Theory and Moral Practice*, *9*(5), 545–561.

Stafford, T. (2014). The perspectival shift: How experiments on unconscious processing don't justify the claims made for them. *Frontiers in Psychology*, *5*, 1067. doi:10.3389/fpsyg.2014.01067

Stewart, B. D., & Payne, B. K. (2008). Bringing Automatic Stereotyping Under Control: Implementation intentions as efficient means of thought control. *Personality and Social Psychology Bulletin*, *34*(10), 1332–1345.

Tetlock, P. F., Kristel, O., Elson, B., Green, M., & Lerner, J. (2000). The psychology of the unthinkable: Taboo trade-offs, forbidden base rates and heretical counterfactuals. *Journal of Personality and Social Psychology*, *78*(5), 853–870.

Turner, M. A., Richardson, T. M., & Ross, S. L. (2007). Housing discrimination in metropolitan America: Unequal treatment of African Americans, Hispanics, Asian Americans, and Native Americans. In J. Goering (Ed.), *Fragile rights within cities: government, housing and fairness*. New York, USA: Rowman & Littlefield Press.

Valian, V. (1999). *Why so slow? The advancement of women*. Cambridge, USA: The MIT Press.

Vargas, M. (2005). The Revisionist's guide to responsibility. *Philosophical Studies*, *125*(3), 399–429.

Vargas, M. (2009). Revisionism about free will: A statement and defense. *Philosophical Studies*, *144*(1), 45–62.

Vargas, M. (2013). *Building better beings: a theory of moral responsibility*. Oxford, UK: Oxford University Press.

Vargas, M. (2017). Implicit bias, responsibility and moral ecology. In D. Shoemaker (Ed.), *Oxford studies in agency and responsibility*. Oxford University Press.

Washington, N., & Kelly, D. (2016). Who's responsible for this? Moral responsibility, externalism and knowledge about implicit bias. In M. Brownstein, & J. Saul (Eds.), *Implicit bias and philosophy* (Vol. 2) (pp. 11–36). Oxford, UK: Oxford University Press.

Watson, G. (1996). Two faces of responsibility. *Philosophical Topics*, *24*(2), 227–248.

Wilson, M. C., & Scior, K. (2014). Attitudes towards individuals with disabilities as measured by the implicit association test: A literature review. *Research in Developmental Disabilities*, *35*(2), 294–321.

WISELI (2012). *Reviewing applicants: research on bias and assumptions* (3rd ed. http://wiseli.engr.wisc.edu/docs/BiasBrochure_3rdEd.pdf accessed 19/09/2016).

Wood, M., Hales, J., Purdon, S., Sejersen, T., & Hayllar, O. (2009). *A test for racial discrimination in recruitment practice in British cities: research report no 607*. London: Department for Work and Pensions.

Zheng, R. (2016). Attributability, accountability and implicit attitudes. In Brownstein, & Saul (Eds.), *Implicit bias and philosophy* (Vol. 2) (pp. 62–89). Oxford, UK: Oxford University Press.

AUTHOR BIOGRAPHIES

Jules Holroyd is a Vice-Chancellor's fellow in Philosophy at The University of Sheffield. Her research focuses on the ways our cognitions may be shaped by and complicit in sustaining broader patterns of injustice. She has published work on implicit bias, on feminist philosophy, and on the moral emotions.

Robin Scaife is a postdoctoral researcher in Philosophy and Psychology at The University of Sheffield. His research focuses on decision making and the degrees to which we have insight and self-control over our decision-making processes.

Tom Stafford is a lecturer in Psychology and Cognitive Science at The University of Sheffield. His research focuses on decision making and skill learning, using measures of behaviour informed by computational, robotics, and neuroscience research.

How to cite this article: Holroyd J, Scaife R, Stafford T. Responsibility for implicit bias. *Philosophy Compass*. 2017;**12**:e12410. http://doi.org/10.1111/phc3.12410