UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Scaife, R., et al. (2020). To Blame? The Effects of Moralized Feedback on Implicit Racial Bias. *Collabra: Psychology,* 6(1): 30. DOI: https://doi.org/10.1525/collabra.251

ORIGINAL RESEARCH REPORT

To Blame? The Effects of Moralized Feedback on Implicit Racial Bias

Robin Scaife, Tom Stafford, Andreas Bunge and Jules Holroyd

Implicit bias training (IBT) is now frequently provided by employers, in order to raise awareness of the problems related to implicit biases, and of how to safeguard against discrimination that may result. However, as Atewologun et al. (2018) have noted, there is very little systematicity in IBT, and there are many unknowns about what constitutes good IBT. One important issue concerns the tone of information provided regarding implicit bias. This paper engages this question, focusing in particular on the observation that much bias training is delivered in exculpatory tone, emphasising that individuals are not to blame for possessing implicit biases. Normative guidance around IBT exhorts practitioners to adopt this strategy (Moss-Racusin et al. 2014). However, existing evidence about the effects of moralized feedback about implicit bias is equivocal (Legault et al. 2011; Czopp et al. 2006). Through a series of studies, culminating in an experiment with a pre-registered analysis plan, we develop a paradigm for evaluating the impact of moralized feedback on participants' implicit racial bias scores. We also conducted exploratory analyses of the impact on their moods, and behavioural intentions. Our results indicated that an exculpatory tone, rather than a blaming or neutral tone, did not make participants less resistant to changing their attitudes and behaviours. In fact, participants in the blame condition had significantly stronger explicit intentions to change future behaviour than those in the 'no feedback' condition (see experiment 3). These results indicate that considerations of efficacy do not support the need for implicit bias feedback to be exculpatory. We tease out the implications of these findings, and directions for future research.

Keywords: Implicit bias; Blame; IAT; Racial bias; Implicit bias training; Moralized feedback

1. Introduction

An action or judgement is implicitly biased when it is influenced by automatic mental processes which distort action or judgement (often without this influence being apparent to the individual). These automatic mental processes are pervasively found: around 70% of the millions who have completed a racial Implicit Association Test (IAT) on the Project Implicit website show an implicit preference for White/light-skin over Black/dark-skin (Nosek et al., 2007). Although recent meta-analyses have revealed a relatively low correlation between IAT results and overt behaviour (Greenwald, Poehlman, Uhlmann, & Banaji, 2009; Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2013), the societal impact of implicit biases has been argued to be significant due to the presence of bias in large numbers of people, and the cumulative effects on individuals of the repeated expression of bias (Greenwald, Banaji, & Nosek, 2015; Buttrick et al., 2020; Kurdi et al., 2019; See also collected authors in Schwlenker 2017 for responses to scepticism about the presence and effects of implicit biases). A large body of research has indicated such processes could influence behaviour in a wide range of contexts. Examples include: differential evaluations of the same CVs where the only difference was race (Dovidio & Gaertner, 2000) — similar effects were found for gender (Uhlmann & Cohen, 2007) and age (Lindner, Graser, & Nosek, 2014) — differential micro-behaviours which display tension and discomfort on the part of White interlocutors, and which inflect the quality of interracial interactions (Dovidio, Gaertner, Kawakami, & Hodson, 2002); and in simulations people mistakenly 'shoot' unarmed Black individuals more frequently than unarmed White individuals —so called 'shooter bias' (Correll, Park, Judd, & Wittenbrink, 2002; Correll, Park, Judd, Wittenbrink, & Sadler, 2007; Plant, & Peruche, 2005).

Importantly, though, in the past decade, research has shown implicit biases are malleable (Lai et al., 2014) and their influence on behaviour can be limited, to some degree. Strategies that have, with varying degrees of success, been shown to mitigate the expression of implicit bias include: attention to counter-stereotypical exemplars (Blair, 2002); the use of implementation intentions to alter patterns of response (Webb, Sheeran, & Pepper, 2010); and the inhibition of automatic associations due to negative affect, such as guilt (Amodio, Devine, & Harmon-Jones, 2007). This suggests implicit biases may be influenced

The University of Sheffield, GB

Corresponding author: Robin Scaife (r.scaife@sheffield.ac.uk)

by other processes, which themselves may be implicit, automatic or otherwise non-conscious. Bias mitigation strategies have been incorporated into models of implicit bias training (IBT) (Devine et al., 2012), although the effects of such strategies on reducing bias have been mixed (Forscher et al., 2017). Some research indicates that merely attributing behaviour to implicit bias rather than explicit bias reduces perceived accountability and makes people less likely to support efforts to combat it (Daumeyer et al., 2019).

One notable feature of many instances of IBT is the exculpatory tone adopted in communicating about implicit bias. Popular implicit bias training resources—like those made public by Facebook, Google and Starbucks—all adopt an exculpatory tone. For example Facebook's 'Managing Unconscious Bias' training (see https://managingbias.fb.com/) describes getting a biased IAT score as indicating that you are a product of the world around you and emphasizes that it is not a comment on who you are as a human being. This is consistent with the recommendation, from Moss-Racusin et al. (2014), that the design of IBT should 'avoid assigning blame' to participants (615). However, the existing evidence about the efficacy of negative or moralised feedback on the manifestation of implicit bias is equivocal. Some studies suggest that negative feedback about implicit bias provokes backlash or hostility (Legault et al., 2011). Other research has indicated interpersonal confrontations may be an effective way of regulating the expression of racial bias (Czopp, Monteith, & Mark, 2006) which could be consistent with an immediate hostile reaction to negative feedback.

In this series of studies, we seek to investigate the impact of tone of feedback (exculpatory, blaming or neutral) on individuals' implicit racial biases, emotions, and intentions to change behaviours. These studies are the first to address this issue. The research goes beyond that of Legault et al. (2011), who focused on external pressures to avoid prejudice, rather than blame and exculpation. The latter evoke moral standards that individuals subscribe to and are motivated to uphold. These studies go beyond the moralised feedback used by Czopp et al. (2006) which was delivered through an online messenger interface. We use specifically moralized feedback (exculpation and blame) delivered in person, in conditions that are good proxies for those of IBT: from an 'authority' figure, to participants who are motivated to avoid prejudice. Some research suggests that activating feelings of guilt might be an effective way of mitigating the expression of implicit bias (Moskowitz & Li 2011), and other research has examined how individuals react to feedback about their IAT scores (Schlachter & Rolf 2017; Howell et al., 2017) or to public discourses about implicit biases in general (Yen, Durrheim & Tafarodi 2018). However, no research has paid attention to the efficacy (or otherwise) of moralised feedback. There is evidence that informing individuals that implicitly biased behaviour is pervasive *increases* the expression of implicit bias (perhaps due to fostering complacency) in the absence of also establishing a strong moral norm against such behaviour (Duguid & Thomas-Hunt, 2015). This suggests setting a strong exculpatory tone may in fact not help with – and may hinder – the mitigation of implicit bias, or the explicit motivation to change.

In sum, we have no direct evidence about whether the standard practice in IBT, of adopting an exculpatory tone, is likely to help or hinder individuals in getting them to modify their implicit and explicit attitudes. Our pre-registered experiment is the first to provide such evidence.

This is also important because wider discussions have, despite equivocal empirical evidence, taken the view that blame is counter-productive and should be avoided. For example, it has been argued blame should be avoided because it is likely to make people less motivated to change (Saul 2013), and that blaming might prevent 'buy in' to norms against implicit bias (Vargas 2017). Further, the idea that implicit bias attributions might let individuals 'off the hook' for their discriminatory behaviour has led some to worry about appeals to implicit biases (Beckles-Raymond 2020). These wider discussions, as well as best practice for IBT, can be informed by evaluating the impact of tone (blaming or exculpatory) on implicit biases, and on intentions to change attitudes and behaviour.

It is important to note that our research focuses on the efficacy (positive or negative effects) of blaming, rather than the warrant (whether it is deserved) for exculpation or blame. The warrant issue is distinct from that of the efficacy of blame or exculpation. Warrant for blame depends on the separate question of responsibility for implicitly biased actions – this issue has been addressed extensively elsewhere (Holroyd, 2012; Holroyd, 2015; Holroyd & Kelly, 2015; Holroyd, Scaife & Stafford, 2017a; Washington & Kelly, 2016), and we set it aside here. Our focus is on efficacy, namely, what the impact is of adopting an exculpatory, neutral, or blaming tone in delivering feedback about implicit bias. Addressing this question does not take on any commitments regarding whether individuals are (solely or collectively) responsible for having or acting on implicit biases.

The importance of developing a detailed understanding of the impact of exculpatory or condemnatory moral communications concerning implicit bias is compounded by the ever-increasing availability of information about our own biases, through online tests such as those at projectimplicit.com, in addition to institutionally provided implicit biases training or reports in the media. This means it is imperative to address the issue of how the tone of such communications impacts upon the expression of implicit bias and readiness to modify implicit and explicit attitudes.

2. Method and Results

2.1. Overview of Research Strategy

We conducted a series of experiments culminating in a pre-registered test of the hypothesis that our blame intervention reduces IAT scores compared to a neutral communication. We also included several exploratory measures to investigate participant's behavioural intentions and self-awareness.

We report how we determined our sample size for the pre-registered experiment, all data exclusions (if any), all

manipulations, and all measures in the study. In addition, we share all experimental materials, files for running the experiments, data and analysis scripts (https://osf.io/awq2c/).

For the purposes of empirically investigating implicit bias we will be equating the phenomenon with performance on indirect attitude measures such as the 'shooter bias' test and the IAT. Focusing on the output of indirect measures does little to define the phenomenon because it does not ensure that any bias measured in this way must always have the features commonly associated with being implicit. Biases measured using the IAT may or may not be outside the agent's awareness, may or may not be aligned with the agent's explicit attitudes, and the picture surrounding the agent's ability to control or alter the bias is certainly unclear (though the responses are required to be fast which inhibits deliberate control). For these reasons we have also included explicit questions regarding the participants' awareness and explicit attitudes in our research to gain further insight. For the purpose of this research we have also assumed that implicit biases must be unendorsed. This is because we are focusing on attitude change and believe that altering a bias that conflicts with explicitly held attitudes is likely to require a different approach to altering a bias that is explicitly endorsed. We resist giving any more detailed definition of implicit bias because, as we have argued in detail elsewhere (see Holroyd, Scaife & Stafford 2017b), there are no unproblematic ways of characterising implicit biases and adopting any definition requires committing to prior theoretical assumptions that we need not take on here.

2.2. Experiment 1

The first experiment was designed to test for a baseline measure of implicit prejudice in our sample and check for any effect of answering explicit attitude questions (to do with moral attitudes) on IAT scores.

2.2.1. Participants

Participants were 83 students from the University of Sheffield (62 female, 21 male; Mean age: 21.66).

Ethnicity: White = 53, Asian/British Asian = 12, Chinese = 11, Mixed, multiple ethnic groups = 4, Arab = 1, Black = 1, Other = 1.

2.2.2. Procedure & Materials

Participants took two IATs; one race: Black/White IAT made up of positive and negative words and images of White and Black young faces cropped at forehead and chin. The other IAT looked for associations with religion (Muslim/non-Muslim) made up of positive and negative words and typical Muslim and Christian biblical names common to a UK population. Implicit Association Tests (IAT) were built using PsychoPy (Peirce, 2007) based on a procedure developed by Greenwald, McGhee, & Schwartz (1998) & Greenwald, Nosek, & Banaji, (2003). The experiment code was published for community use as 'OpenIAT' (See https://osf.io/i8yj5/ Stafford & Scaife, 2014). The race version used images from Project Implicit stimulus materials (Nosek et al., 2007). The religion version used typical Muslim and Christian biblical names. See supplementary information for details of both IATs.

Participants also answered a number of explicit attitude questions covering their attitudes towards: the racial and religious groups which were the focus of the IATs, racial and religious integration, and moral norms concerning the equal treatment of racial and religious groups (see supplementary information). In order to allow us to check for any influence on IAT scores caused by considering the issues raised in the explicit questions the order was counter-balanced so half the participants answered the explicit attitude questions before taking the two IATs and half after taking the two IATs.

2.2.3. Results

The IAT results were commensurate with those found in the wider literature. As shown in **Figure 1**, the participants had a significantly non-zero implicit preference for White over Black faces (mean race IAT score: +0.48, SD 0.56) $t(82) = 7.88$, $p < 0.000001$) and a slightly smaller preference for non-Muslim over Muslim names (mean religion +0.37 SD 0.58 $t(81) = -5.88$, $p < 0.000001$). This provides an important demonstration that implicit anti-Black and anti-Muslim biases are replicated when IATs are conducted using a UK sample. The explicit attitude questions indicated that participants had no explicit racial or religious preferences and that they believed it unacceptable to judge people on these social identities.

An analysis of the impact of answering the explicit questions before taking the IATs (mean = +0.43, SD 0.47) rather than after (mean = +0.41, SD 0.38) indicated that there was no significant order effect $t(81) = 0.22576$, $p = 0.822$. Effect Size 0.02 [CI –0.40 +0.46] (this and all effect sizes reported in this paper are Cohen's d).

An equivalence test (Lakens, 2017; Lakens, Scheel & Isager, 2018) was conducted. The equivalence test was non-significant, $t(78.64) = 1.590$, $p = 0.0579$, given equivalence bounds of –0.171 and 0.171 (on a raw scale) and an alpha of 0.05. The equivalence test bounds were based on a medium sized effect of 0.4 d. This suggests the observed effect is statistically not different from zero and statistically not equivalent to zero. This indicates that considering moral judgements about race and ethnicity did not impact on the expression of implicit bias on either IAT.

2.3. Experiment 2

The second experiment was designed to test the impact of two interpersonal moral communications on implicit bias scores. The moral communication of blame was anticipated to activate guilt, which may reduce the expression of implicit bias (Moskowitz & Li, 2011). The focus on a race IAT was partly because experiment 1 found more prejudiced scores on the race IAT than on the religion IAT and partly because of complexities in identifying what social identity was tracked by the religion IAT (participants' biases that target Muslim identity may include racialized components as well as assumptions about geographical origins, religious or doctrinal commitments).

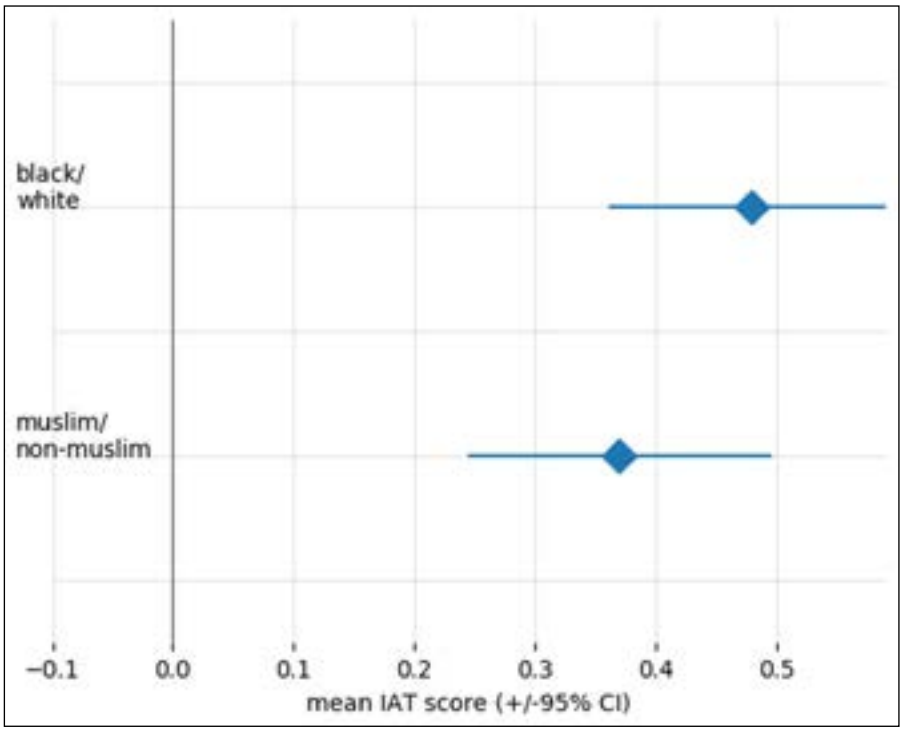

Figure 1: Scores on two IATs for 83 participants in experiment 1.

2.3.1. Participants

Participants were 121 students from the University of Sheffield. 5 participants were excluded from all the experiment 2 analyses either because their understanding of English was poor or they did not engage with the task properly. This left 116 participants (87 female, 29 male; Mean age: 19.66). Ethnicity: White = 91 East Asian = 18 South Asian= 5 Black = 1 Mixed = 1.

2.3.2. Materials

Race IAT as used in experiment 1.

'Shooter bias' test built using PsychoPy (Peirce, 2007) based on the task and using the images from 'The police officer's dilemma' (Correll, Park, Judd, & Wittenbrink, 2002). This task was selected precisely because the behaviour engaged (shooting, albeit simulated) is liable to moral evaluation and so makes moralized feedback (blame or exculpation) appropriate.

Explicit attitude questions, including moral attitudes: see supplementary information.

2.3.3. Procedure

Having completed the information and consent form (approved by the Department of Psychology, University of Sheffield) participants were seated at a laptop in a small room with the door open to a larger room where the experimenter was based at a computer desk. The experimenter gave participants a brief introduction to the study in which they were told that after the first task was finished they would be provided with feedback on their performance before they could move onto the second part of the experiment. Participants then undertook the shooter bias test. The instructions informed them to 'shoot' armed targets and to 'not shoot' unarmed targets. They were also told they have less than a second (0.85s) to make each decision. Targets were Black and White males who appeared on complex backgrounds either holding guns or non-weapon objects such as phones or drinks cans. The task was made up of 80 trials (20 Black armed, 20 Black unarmed, 20 White armed, 20 White unarmed) presented in a random order. After each choice, the participant was presented with feedback text on the screen indicating either: 'Correct choice' or 'Error'. If the participant took longer than 0.85 of a second to respond the trial timed out and the 'too slow' feedback was presented. Once the shooter bias task was over the experimenter gave the appearance of running some analysis code and looking at a graph supposedly representing the participant's responses. Following this the experimenter delivered one of two types of 'feedback' depending on which condition the participant was in. Participants in the blame condition where told:

> 'You have just taken the shooter bias test, which is intended to measure differences in attitudes towards racial groups that you might not explicitly endorse. I'm afraid that the differences in your reaction times and shooting choices indicate you have negative implicit attitudes towards Black people. Morally speaking, we would hope people don't have these kinds of attitudes. People who have these kinds of attitudes tend to behave in discriminatory ways, even if it is so subtle that you don't notice it. Overall, you are blameworthy for having these discriminatory attitudes and behaviours. As you probably know, it is morally unacceptable to have biased attitudes and behaviours; it would be quite normal to feel guilty about this; and to think about how to

> change these attitudes, or your behaviours to bring them in line with moral expectations. Later, in the debrief, we can talk more about techniques people have used to try to eliminate these bad attitudes. There'll also be the chance to ask any questions you may have. Now that you've got the results of this part of the study, we'll give you a moment to reflect on that, and then move on to the next part of the study.'

This communication included key components of a variety of views on the nature of blame: it included the expression of a shared moral standard to which the individual is expected to adhere, and has fallen short ('we would hope you don't have these attitudes'; 'it is morally unacceptable to have biased attitudes and behaviours'); expression of emotional response ('I'm afraid that…'); anticipated emotive responses ('it would be normal to feel guilty about this'); and reaffirmation of the moral standard ('think about how to change these attitudes, or your behaviours …') (Wallace, 1994; Strawson, 1962; Bennett, 2002; Fricker, 2014). It is important to note the text above is not supposed to capture a 'folk' conception of blame; rather to present a theoretical construct that may overlap with a variety of different conceptions of blame. The manipulation aimed to incorporate different components of various views on blame, capturing all components that may be critical for the purposes of efficacy, whilst avoiding construing blame so narrowly as to misalign with any one participant's understanding of blame. So, for example, we did not include reference to the idea that implicit bias is due to the culture you live in, although this is a common trope, since this claim would provide a basis for participants to contest the blame. Moreover, the blame was communicated in a low emotional tone, in accordance with the finding that individuals most readily recognised such responses as blaming ones (Malle, Guglielmo, & Monroe, 2014). The blaming communication is not supposed to be universal, but to capture the components of blame that we expected to resonate with UK populations. It may be that a different blame manipulation would be required for a different audience with different background assumptions about moral communications. Participants in the exculpation condition where given the same feedback except that rather than being told they are blameworthy and it would be normal to feel guilty, they were informed that they are entirely blameless, not culpable, and should not feel guilty (see full exculpation script below).

> 'You have just taken the shooter bias test, which is intended to measure differences in attitudes towards racial groups that you might not explicitly endorse. I'm afraid that the differences in your reaction times and shooting choices indicate that you have negative implicit attitudes towards Black people. Morally speaking, we would hope people don't have these kinds of attitudes, but it is actually entirely blameless. People who have these attitudes tend to behave in discriminatory ways, but it is so subtle you don't notice it, and you aren't culpable for doing so. It would be quite easy to feel guilty about these attitudes and behaviours, but you shouldn't. In fact, it would be great if you were concerned to consider the various steps that can be taken to change these attitudes, or your behaviours. Later, in the debrief, we can talk more about techniques people have used to try to eliminate these attitudes. There'll also be the chance to ask any questions you may have. Now that you've got the results of this part of the study, we'll give you a moment to reflect on that, and then move on to the next part of the study.'

Some participants interrupted their feedback with comments (typically trying to either explain or deny their bias), or with questions (typically about the test or how common implicit bias is). When this occurred, they were told 'There will be a chance for you to make comments and ask questions in the debrief at the end. For now please continue with the experiment.' After receiving their 'feedback' participants answered a number of explicit attitude questions from experiment 1 (see supplementary information) and undertook the race IAT from experiment 1. At the end of the experiment participants also took part in a task where they were asked to rate the suitability of Black or White job candidates. This was designed to evaluate the impact of our intervention on behavioural tasks. However, the task did not yield any significant results, partly because the participants seemed to be overly keen to demonstrate they were not racist by over-evaluating the suitability of Black candidates (see report on additional measures for more details at https://osf.io/sjz5t/). This indicates a moral or social norm to avoid anti-Black racism was clearly evoked and felt by the participants. After the experiment participants were thoroughly debriefed regarding the scripted nature of their feedback.

2.3.4. Results

There was no significant difference between the IAT scores of participants in the blame (mean +0.18, SD 0.78) and exculpation (mean +0.24, SD 0.67) manipulations $t(113.32) = 0.37629$, $p = 0.7074$, effect size $d = 0.035$ [CI –0.33 +0.40]. Both means are considerably lower than the baseline scores from experiment 1 (see **Figure 2**).

An equivalence test (Lakens, 2017; Lakens, Scheel & Isager, 2018) was conducted. The equivalence test was significant, $t(113.32) = 1.781$, $p = 0.0388$, given equivalence bounds of –0.292 and 0.292 (on a raw scale) and an alpha of 0.05. The equivalence test bounds were based on a medium sized effect of 0.4 d. This suggests that the observed effect is statistically not different from zero and statistically equivalent to zero.

The explicit questions indicated many participants in the exculpation manipulation reported feeling blamed (23 out of 56) and/or guilty (49 out of 56). This suggests the exculpation manipulation was not successful in producing feelings of exculpation. The attribution of self-blame, or feelings of guilt – irrespective of whether blame or exculpation was communicated by the experimenter – may then have driven the effects. This would mean that

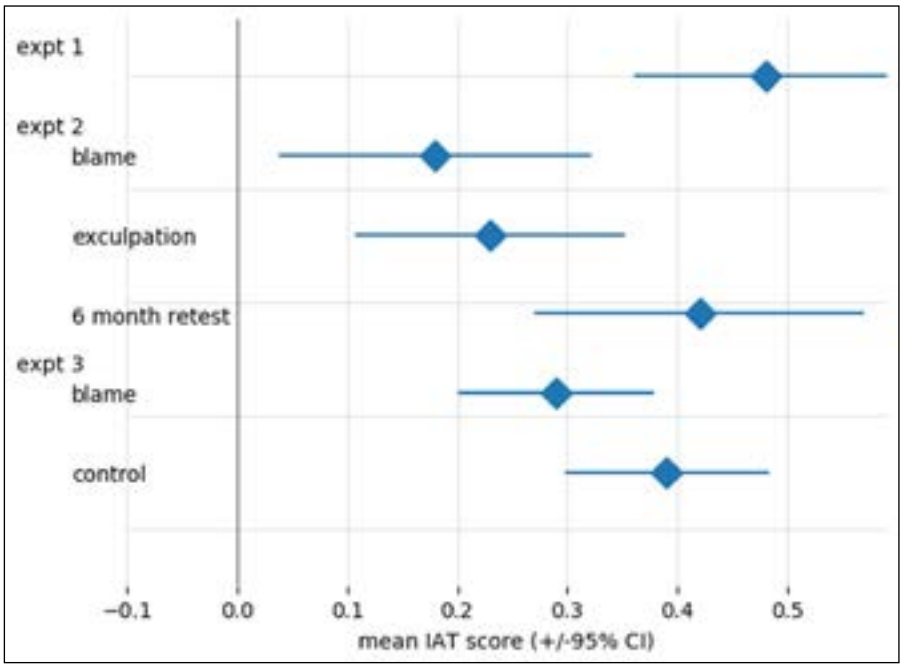

Figure 2: Mean IAT scores on race IAT by experiment & condition.

the exculpation condition would not be a valid test of the effects of actually removing blame or guilt (though it may be a good test of attempts to remove blame or guilt, if people generally tend to feel guilt in the face of exculpatory messages).

Explicit attitude questions confirmed that whilst demonstrating a non-zero IAT, nearly all our participants indicated that it is both socially and morally unacceptable to make judgements about people based on their race.

2.4. Six month follow up

2.4.1. Participants

Participants were 60 students from the University of Sheffield all of whom had taken part in experiment 2. Participants were from both conditions of experiment 2 (36 from the blame condition, 24 from the exculpation condition). 46 Female, 14 Male. Mean age 19.67. Ethnicity: White = 50 East Asian = 9 South Asian = 1.

2.4.2. Materials

Race IAT as used in experiment 1.

Explicit attitude questions: see supplementary information.

2.4.3. Procedure

Participants were recruited via email invitation. Participants took our race IAT and answered a number of explicit questions about their experience of experiment 2 and how it had affected them since (see supplementary information).

2.4.4. Results

We found participants' IAT scores (mean +0.42, SD 0.59) had returned to levels commensurate with participants from experiment 1 who had not engaged in a moral communication (mean +0.48, SD 0.56). The IAT showed only a weak test re-test reliability (Pearson $r = 0.26$ ($p = 0.047$), Spearman's rank correlation = 0.29 ($p = 0.023$)); but it should be noted this is not a pure assessment of test-retest reliability, since participants at time 1 were in one of the two moral communication conditions (blame or exculpation). Note the wider literature estimates test-retest reliability of the IAT as modest -typically between $r = .5$ or $r = .6$ (Nosek 2007, Greenwald et al., 2015) – although the interval used for those retests is substantially shorter than the 6 month gap between tests in our research.

We also asked participants about their experience of taking part in experiment 2. The highest emotion ratings on a 1 to 7 scale were for feeling guilty (mean 5.12, SD 1.6) and being upset (mean 4.4, SD 1.68). Full details of participant's emotional ratings can be found in the supplementary information. Despite reporting these strong negative emotional responses over 90% of participants said they were glad they took part in the experiment. 68% of participants reported that taking part in the experiment had made them less likely to make prejudiced judgments and 68% reported that they had done something to try to ensure they treat all people equally since taking part in the initial experiment (note that these two 68% groups are made up of different but mostly overlapping participants). No participants indicated strong regret about taking part and only 3% expressed mild regret about having taken part. We note that selective recruitment is likely to bias these results (i.e., that participants with higher regrets would be less likely to participate in follow up testing).

2.5. Experiment 3

Experiment 3 was pre-registered (https://osf.io/94pur/) to confirm a difference on IAT scores when participants experience blame for the implicit associations (the

blame condition from experiment 2) compared to a no blame control (similar to the no feedback condition of experiment 1). One reason for conducting this study was to test if the difference persisted when the control group was specifically designed to mimic the experimental condition in all respects except the blame feedback. This allowed us to rule out that factors such as taking the shooter bias test caused practice effects or contributed towards the observed difference between the two groups in any other way. Another reason for focusing on blame rather than exculpation is that experiment 2 showed it was hard to design an exculpatory communication which didn't provoke feelings and reactions similar to those that result from being blamed. Thus, it is important to confirm whether responses to blame are counter-productive or constructive, compared to a neutral communication. Statistical power analysis determined that a sample of 160 would be adequate, assuming a minimum effect size of interest of 0.4 and 80% power). This minimum effect size was based on analysis comparing the IAT scores of participants in experiment 1 to scores of those in the blame condition from experiment 2, which showed a standardised difference of means of 0.45. As well as the outcome variable supporting the pre-registered comparison (IAT scores after the blame vs no feedback manipulation), a number of other variables were included for exploratory analysis. These include explicit behavioural intentions (building on the findings from experiment 2b), self-awareness of bias (as previously investigated by Hahn, Judd, Hirsh, & Blair, 2014) and intellectual humility (as previously investigated by McElroy et al., 2014). We report these here, insofar as they directly complement understanding of how blame impacted on participant's attitudes, and report fully all measures and data in the online supplementary material. We note that the pre-registered component of experiment 3 refers only to the effect of the manipulation on the primary outcome variable, namely, IAT scores after the manipulation.

2.5.1. Participants

All 162 participants were students from the University of Sheffield (100 female, 61 male, 1 agender). Mean age 20.34.

Ethnicity: White (British/White British/Caucasian/English/European/Greek Cypriot): 134 (83%), South Asian (British Asian/Asian Indian/Pakistani/Indian/South Asian): 9 (6%), East Asian (Asian/Chinese/Vietnamese/Burmese): 10 (6%), Mixed: 5 (3%), White Asian: 2 (1%), Black (Black British/Caribbean): 2 (1%).

No participants were completely excluded from the analysis but 3 participants were excluded from most of the analysis (2 for demonstrating an awareness of our hypothesis in the post-experimental debriefing and 1 for indicating racial preferences on the explicit attitude questions).

2.5.2. Materials

Race IAT as used in experiment 1 and 2.

Shooter bias test as used in experiment 2.

Explicit attitude questions: see supplementary information.

Intellectual Humility Scale: 13-item scale developed by McElroy et al. (2014).

2.5.3. Procedure

The method was as in experiment 2 where all participants took the shooter bias test before receiving feedback except the exculpation condition was replaced by a control condition in which participants received a scripted response with no information about their performance and no moral evaluation. This was intended to provide an interpersonal communication without promulgating feelings of guilt in control participants, which were not eliminated by the exculpation communication used in experiment 2. Participants in the control condition were told the following:

> 'You have just taken the first test, which is intended to measure your reactions to different stimuli that you might not be familiar with. I've just checked your data has come through OK. It won't tell us anything meaningful until we have all the results in. Later, in the debrief, we can talk more about the nature of our research if you like. There'll also be the chance to ask any questions you may have. Now that we've done this part of the study, we'll give you a moment to rest, and then move on to the next part of the study.'

This 'no feedback' condition is a useful proxy for the context of IBT sessions in which participants are told about the phenomenon of implicit bias, but do not receive individualized feedback about IAT scores. As in experiment 2, after receiving their feedback participants answered a number of explicit attitude questions and undertook the race IAT. Following the IAT, experiment 3 also included additional explicit questions to access participant's future behavioural intentions, awareness of their own level of bias, emotional responses, and an intellectual humility scale. After this but before the debrief participants also completed a seating distance measure and a voluntary time commitment assessment which were intended to monitor the potential behavioural impact of the intervention.

2.5.4. Results

The raw data and analysis scripts are available at: https://osf.io/awq2c/.

2.5.4.1. Pre-registered test of effect on IAT scores analysis

The mean IAT score for the blame group was +0.29 (SD 0.58). The control mean was +0.39 (SD 0.60). The difference between these groups was not significant $t(157) = 1.11$, $p = 0.135$ (Cohen's d effect size = 0.18, 95% confidence interval –0.14, 0.49). These confidence intervals suggest the true effect of blaming individuals is more likely to be neutral or a reduction in implicit prejudice. The possibility that blame *increases* implicit prejudice by any non-small amount is actively precluded. **Figure 2** below shows the variation in mean race IAT scores across all experiments and conditions, showing the

extent to which blame and exculpation manipulations reduce negative (anti-Black) implicit bias scores.

2.5.4.2. Exploratory Analysis

An equivalence test (Lakens, 2017; Lakens, Scheel & Isager, 2018) was non-significant (t(159.82) = −1.467, p = 0.0722, given equivalence bounds of +/− 0.4 d, +/− 0.236 on the raw scale, and an alpha of 0.05. Equivalence bounds based on the effect size of 0.4, used in the pre-registration). This suggests that, although the IAT scores were not significantly different between conditions there was also insufficient evidence to conclude they are equivalent.

Behavioural Intentions:

Participants were asked: 'Do you intend to try to change your future behaviour as a result of your experience in this experiment?' This question provided evidence that the blame manipulation has a positive impact on people's explicit intentions to change their future behaviour (mean 5.03, SD 1.77) when compared to the behavioural intentions of the control group (mean 3.48, SD 1.86), t(157) = 5.38, $p < 0.000001$. Effect size: d = 0.85. This is so despite the fact that the communication of blame produced an atmosphere that was intense and uncomfortable for the participant, producing a significant increase in anxiety (blamed mean 3.46, SD 1.72, control mean 2.77 SD 1.67; t(160) = 2.60, p = 0.01. Effect size: d = 0.41). Furthermore, the degree of intended behaviour change correlates positively with the degree to which participants felt blamed r(160) = 0.22, p = 0.0048. It also correlates positively with the degree to which participants felt guilty r(160) = 0.17, p = 0.03.

3. Discussion

This research is the first to investigate the effects of blame on individuals' implicit biases (with a planned, pre-registered, comparison) and the effect on intentions to change behaviours. This was done using a moral communication produced by a rich, 'in person' dialogue that could serve as a proxy for communication in the context of IBT. The results contradict the view that blaming people for implicit bias is a counterproductive approach to addressing implicit biases, either by increasing implicit bias, making people less motivated to change or preventing 'buy in' to the project of counteracting implicit bias. The findings suggest blaming does not increase implicit bias, as measured by the IAT, at least by any significant amount. The results are compatible with blame reducing IAT scores, but the current experiments were not sufficiently powered to distinguish between a zero and positive non-zero effect on IAT scores. Further research is needed to differentiate between these two conclusions. Exculpating communications were experienced as guilt provoking (experiment 2), but there was also no evidence these increased implicit racial biases (see **Figure 2**). Second, blaming responses were correlated with stronger intentions, as compared to no feedback, to change future behaviours to combat implicit bias and racial discrimination. This shows that assumptions about blame being problematic may be mistaken, both in relation to the impact on implicit biases, and explicit intentions to change.

Whilst our findings show blame does not itself reduce implicit bias, they do suggest that communications which are taken to have a moral flavour (such as the provision of moralised feedback on one's personal implicit biases) can be important in motivating individuals to form explicit intentions to change behaviours influenced by implicit racial biases. Whilst our explicit emotional measures indicate that even the non-confrontational blame feedback caused a significant increase in anxiety, so did the communication of an exculpatory message. Moreover, these negative emotional reactions did not impact negatively on the reported intention to change behaviour. In fact, individuals who received the blaming manipulations showed the *strongest* expression of such intentions. This is consistent with earlier findings that guilt is efficacious in motivating commitments to behavioural change. Further analyses of the relative efficacy of communicating negative but non-moralised feedback versus negative but moralised feedback is required. It is worth noting that the 'feedback' involved in the moral communication tested in these studies was bogus feedback – it was not based on individual's actual implicit bias measures, but consistently communicated that all participants expressed some anti-Black implicit bias (although our studies, in line with other work on implicit biases, suggest that this invariant feedback would have been accurate for most participants in ascribing them non-zero implicit bias).

Most interventions do not bring about long-term change in implicit bias (Lai et al., 2016). This may be true of our interventions in experiment 2, since a retest on 50% of the participants 6 months later demonstrated that participants' IAT scores returned to levels commensurate with those of participants from experiment 1 who had not engaged with a moral communication (although strong conclusions are precluded due to selective drop-out from our follow up test). Accordingly, the finding that blame had an impact on individuals' explicit intentions to change their behaviour is particularly important, since this may be the most significant driver of change. It is of course possible that the self-reported changes in explicit intentions are due to social desirability effects rather than individuals internalising the relevant moral norms. But this, after all, is a legitimate purpose of moralized feedback (to change social norms, due to either external or internal pressure). Note, moreover, that to the extent that social pressure explains the uniform commitment to anti-racist values in explicit questions, there is not the same uniformity in expressions of intentions to change attitudes and behaviour, suggesting that social pressure alone cannot account for the latter effect. Future studies could be devised to differentiate between these two explanations. It is however worth noting that even a shift caused by desirability effects could still be of great value, particularly if it can cause individuals to act to reduce their prejudice.

Another alternative interpretation of our findings, in the context of these remarks about guilt, is an 'Aversive-Arousal

Reduction Hypothesis' akin to the one considered by Batson (1991). This hypothesis explains individuals' expressed commitments to change behaviour in terms of assuaging the social anxiety or discomfort that was produced by the moral communications. This explanation would undermine the proposal that blame is instrumental in producing constructive explicit attitudes, since the motivation to change would not really be about reducing racism or combating implicit biases; and the resolution to change would not outlive its role in reducing anxiety once the experience of taking part in the experiment is over. However, this interpretation is not supported by our findings. First, there is evidence that the motivation to change did indeed persist (from our six-month follow up of experiment 2). Moreover, when individuals were given an opportunity to escape social anxiety or feelings of guilt — those who were in the exculpation condition in experiment 2 — they did not take this opportunity; many individuals in the exculpation condition reported feeling self-blame and guilt. Accordingly, this alternative interpretation is not defensible on the basis of our findings.

Whilst the blame intervention in experiment 3 did not significantly decrease implicit bias scores, the results of our behavioural intention measure do provide a strong indication that blaming individuals has a positive impact on their explicit intentions to reduce their prejudice. Moreover, we found this indication on the basis of a one-shot intervention: the cumulative effects of such moral communications, or of a pervasive moral norm against being influenced by implicit racial bias, may be even more powerful in terms of long-term changes in awareness, stronger or more motivationally effective moral emotions, and greater commitments to attitudinal and behavioural change.

One caveat for generalising this conclusion may be that our sample was made up of young people in higher education, who had indicated on an explicit attitude question that it was unacceptable to treat people differently based on race. This sample may be more receptive to feedback than the general population. However, where IBT is not mandatory, one would expect those who volunteer for such training to be similarly receptive to feedback, so this feature of our sample helpfully aligns with expected real-world scenarios in which tone of communication about implicit bias is important. Indeed, Howell et al. (2017) found participants assigned to take an IAT were more defensive in response to feedback indicating they have biased attitudes than those who self-selected to take part on the project implicit website. This defensiveness centred on the validity of the IAT or how accurately it reflected their attitudes. Those with this defensiveness were unlikely to intend to change their behaviour. In contrast those who felt the worst about their feedback were most willing to engage in bias reducing strategies. This supports the idea that blame causing negative affect may help motivate bias reducing behaviour change. However, a certain level of initial receptiveness may be required to ensure that the feedback is taken seriously.

Moreover, further studies are needed to identify which features of blame increase participant's explicit intentions to change their behaviour and if these intentions have a significant impact on their actual future behaviour — including, for example, their engagement with other, more effective implicit bias reduction strategies. Further limitations of the present research are that it only investigated one type of blaming communication, where blame was delivered in a low emotion non-confrontational manner by an unfamiliar member of psychology staff whom the participants are likely to view as an authority figure. However, again, this is likely to approximate the conditions in which feedback is delivered in IBT. Moreover, our study focused on a case in which there was high buy-in to the moral norm violated, as indicated by the self-report measures of explicit attitudes. This buy-in was recognised and emphasised in the moral communication, which made reference to the pervasive norm against racial bias (recall the wording: 'as you probably know, it is morally unacceptable…'; 'it would be quite normal to feel guilty…'; changing behaviours 'to bring them into line with moral expectations…'). It may be that blame delivered in a more confrontational manner, by someone who stands in a different relationship to the individual being blamed, would produce different effects. Likewise, blame for the violation of a more controversial norm, or blame for violation of a norm that the individual does not endorse, may produce different effects. These possibilities could range from more significant decrease in bias (perhaps for more confrontational feedback) or even a rebound increase in bias (perhaps for violation of controversial norms). It could also be the case that the precise nature of the moral communication does not cause any significant differences. These are areas for further research.

Finally, caution is required in understanding the practical implications of these results, and their relevance to current institutional practice. We noted that institutional training sessions tend to set an exculpatory tone, and that this is recommended 'best practice'. Our findings challenge the assumption that anything but an exculpatory tone is counter-productive as we found no evidence that blame causes backlash. Whilst implicit racial biases did not significantly decrease, nor did they increase following blame. In fact, our behavioural intention measure provided evidence that blame seems efficacious in encouraging the explicit motivation to change behaviour. But that blame is efficacious in at least the formation of explicit intentions does not entail that we *should* blame. Exculpatory communications were experienced as producing guilt and self-blame, but appeared to be less effective in reducing bias or promoting intentions to change behaviour. Whether communications of blame or exculpation are all things considered justified may depend on further considerations, including questions concerning individual or collective moral responsibility for possessing and manifesting implicit bias (namely, the warrant for blame), and on the propriety of blaming responses — which may depend on the relationship of the blamer to the blamed. However, the current investigation suggests there are no reasons grounded in considerations of anticipated backlash to refrain from blaming (though

there may be others). Nor do they vindicate exculpation as the obviously efficacious strategy. This is the first set of experimental studies, including a pre-registered component, to investigate the important question of tone of delivery of feedback about implicit bias. The evidence from these studies points to the need for more future research, and supports critical reflection on the common practice of delivering implicit bias training in an exculpatory manner.

Data Accessibility Statement

All the stimuli, presentation materials, participant data, and analysis scripts can be found on this paper's project page on the Open Science Framework https://osf.io/awq2c/.

Additional File

The additional file for this article can be found as follows:

- **Text S1.** Additional details of experiments. DOI: https://doi.org/10.1525/collabra.251.s1

Acknowledgements

Thanks for comments and advice are due to Dan Kelly and Paschal Sheeran.

Funding Information

This work was supported by a Leverhulme Trust Project grant 'Bias and Blame' project [grant number: RPG- 2013-326].

Competing Interests

JH, TS & RS have received payment in the past for delivering lectures and/or training on implicit bias.

Author Contributions

- Contributed to conception and design: JH, TS, AB & RS
- Contributed to acquisition of data: TS & RS
- Contributed to analysis and interpretation of data: JH, TS, RS
- Drafted and/or revised the article: JH, TS, AB, RS
- Approved the submitted version for publication: JH, TS, AB, & RS

References

Amodio, D., Devine, P., & **Harmon-Jones, E.** (2007). A Dynamic Model of Guilt: Implications for Motivation and Self-Regulation in the Context of Prejudice. *Psychological Science*, *18*(6), 524–530. DOI: https://doi.org/10.1111/j.1467-9280.2007.01933.x

Atewologun, D., Cornish, T., & **Tresh, F.** Unconscious bias training: An assessment of the evidence for effectiveness. *EHRC Research Report*, 2018. https://www.equalityhumanrights.com/en/publication-download/unconscious-bias-training-assessment-evidence-effectiveness

Batson, C. D. (1991). The Altruism Question: Toward a Social-Psychological Answer. Hillsdale, NJ: Lawrence Erlbaum Associates. DOI: https://doi.org/10.4324/9781315808048

Beckles-Raymond, G. (2020). Implicit Bias, (Global) White Ignorance, and Bad Faith: The Problem of Whiteness and Anti-Black Racism. *Journal of Applied Philosophy*. https://onlinelibrary.wiley.com/doi/full/10.1111/japp.12385. DOI: https://doi.org/10.1111/japp.12385

Bennett, C. (2002). The varieties of retributive experience. *The Philosophical Quarterly*, *52*(207), 145–163. DOI: https://doi.org/10.1111/1467-9213.00259

Blair, I. (2002). The Malleability of Automatic Stereotypes and Prejudice. *Personality and Social Psychology Review*, *6*(3), 242–261. DOI: https://doi.org/10.1207/S15327957PSPR0603_8

Buttrick, N., Axt, J., Ebersole, C. R., & **Huband, J.** (2020). Re-assessing the incremental predictive validity of Implicit Association Tests. *Journal of Experimental Social Psychology*, *88*, 103941. DOI: https://doi.org/10.1016/j.jesp.2019.103941

Correll, J., Park, B., Judd, C. M., & **Wittenbrink, B.** (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of personality and social psychology*, *83*(6), 1314–1329. DOI: https://doi.org/10.1037/0022-3514.83.6.1314

Correll, J., Park, B., Judd, C. M., Wittenbrink, B., & **Sadler, M. S.** (2007). Across the thin blue line: Police officers and racial bias in the decision to shoot. *Journal of Personality and Social Psychology*, *92*(6), 1006–1023. DOI: https://doi.org/10.1037/0022-3514.92.6.1006

Czopp, A. M., Monteith, M. J., & **Mark, A. Y.** (2006). Standing up for a change: Reducing bias through interpersonal confrontation. *Journal of personality and social psychology*, *90*(5), 784–803. DOI: https://doi.org/10.1037/0022-3514.90.5.784

Daumeyer, N. M., Onyeador, I. N., Brown, X., & **Richeson, J.** (2019). Consequences of Attributing Discrimination to Implicit vs. Explicit Bias. DOI: https://doi.org/10.31234/osf.io/42j7v

Devine, P. G., et al. (2012). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of experimental social psychology*, *48*(6), 1267–1278. DOI: https://doi.org/10.1016/j.jesp.2012.06.003

Dovidio, J. F., & **Gaertner, S. L.** (2000). Aversive racism and selection decisions: 1989 and 1999. *Psychological Science*, *11*, 319–323. DOI: https://doi.org/10.1111/1467-9280.00262

Dovidio, J. F., Gaertner, S. L., 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), 88–102. DOI: https://doi.org/10.1037/1099-9809.8.2.88

Duguid, M. M., & **Thomas-Hunt, M. C.** (2015). Condoning stereotyping? How awareness of stereotyping prevalence impacts expression of stereotypes. *The journal of Applied Psychology*, *100*(2), 343–59. DOI: https://doi.org/10.1037/a0037908

Forscher, P. S., et al. (2017). Breaking the prejudice habit: Mechanisms, timecourse, and longevity. *Journal of experimental social psychology*, *72*, 133–146. DOI: https://doi.org/10.1016/j.jesp.2017.04.009

Fricker, M. (2014). What's the Point of Blame? A Paradigm Based Explanation. *Noûs*, *50*(1), 165–183. DOI: https://doi.org/10.1111/nous.12067

Greenwald, A. G., Banaji, M. R., & **Nosek, B. A.** (2015). Statistically small effects of the Implicit Association Test can have societally large effects. *Journal of Personality and Social Psychology*, *108*, 553–561. DOI: https://doi.org/10.1037/pspa0000016

Greenwald, A. G., McGhee, D. E., & **Schwartz, J. L. K.** (1998). Measuring Individual Differences in Implicit Cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480. DOI: https://doi.org/10.1037/0022-3514.74.6.1464

Greenwald, A. G., Nosek, B. A., & **Banaji, M. R.** (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*(2), 197–216. DOI: https://doi.org/10.1037/0022-3514.85.2.197

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & **Banaji, M. R.** (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of personality and social psychology*, *97*(1), 17–41. DOI: https://doi.org/10.1037/a0015575

Hahn, A., Judd, C. M., Hirsh, H. K., & **Blair, I. V.** (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology: General*, *143*(3), 1369–1392. DOI: https://doi.org/10.1037/a0035028

Holroyd, J. (2012). Responsibility for implicit bias. *Journal of Social Philosophy*, *43*(3), 274–306. DOI: https://doi.org/10.1111/j.1467-9833.2012.01565.x

Holroyd, J. (2015). Implicit bias, awareness and imperfect cognitions. *Consciousness and cognition*, *33*, 511–523. DOI: https://doi.org/10.1016/j.concog.2014.08.024

Holroyd, J., & **Kelly, D.** (2015). Implicit bias, character and control. In A. Masala & J. Webber (Eds.), From Personality to Virtue, Oxford University Press, 106–133. DOI: https://doi.org/10.1093/acprof:oso/9780198746812.003.0006

Holroyd, J., Scaife, R., & **Stafford, T.** (2017a). Responsibility for implicit bias. *Philosophy Compass*, *12*(3), e12410. DOI: https://doi.org/10.1111/phc3.12410

Holroyd, J., Scaife, R., & **Stafford, T.** (2017b). What is implicit bias? *Philosophy Compass*, *12*(10), e12437. DOI: https://doi.org/10.1111/phc3.12437

Howell, J., Redford, L., Pogge, G., & **Ratliff, K.** (2017). Defensive Responding to IAT Feedback. *Social Cognition*, *35*, 520–562. DOI: https://doi.org/10.1521/soco.2017.35.5.520

Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., ..., & **Banaji, M. R.** (2019). Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *The American psychologist*, *74*(5), 569–586. DOI: https://doi.org/10.1037/amp0000364

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J. L., Joy-Gaba, J. A., Ho, A. K., Teachman, B. A., Wojcik, S. P., Koleva, S. P., Frazier, R. S., Heiphetz, L., Chen, E., Turner, R. N., Haidt, J., Kesebir, S., Hawkins, C. B., Schaefer, H. S., Rubichi, S., Sartori, G., Dial, C. M., Sriram, N., Banaji, M. R., & **Nosek, B. A.** (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General*, *143*, 1765–1785. DOI: https://doi.org/10.1037/a0036260

Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., Calanchini, J., Xiao, Y. J., Pedram, C., Marshburn, C. K., Simon, S., Blanchar, J. C., Joy-Gaba, J. A., Conway, J., Redford, L., Klein, R. A., Roussos, G., Schellhaas, F. M. H., Burns, M., Hu, X., McLean, M. C., Axt, J. R., Asgari, S., Schmidt, K., Rubinstein, R., Marini, M., Rubichi, S., Shin, J-E. L., & **Nosek, B. A.** (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General*, *145*(8), 1001–1016. DOI: https://doi.org/10.1037/xge0000179

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social psychological and personality science*, *8*(4), 355–362. DOI: https://doi.org/10.1177/1948550617697177

Lakens, D., Scheel, A. M., & **Isager, P. M.** (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. DOI: https://doi.org/10.1177/2515245918770963

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. DOI: https://doi.org/10.1177/0956797611427918

Lindner, N. M., Graser, A., & **Nosek, B. A.** (2014). Age-based hiring discrimination as a function of equity norms and self-perceived objectivity. *PLoS ONE*, *9*, 1–6. DOI: https://doi.org/10.1371/journal.pone.0084752

Malle, B. F., Guglielmo, S., & **Monroe, A. E.** (2014). A theory of blame. *Psychological Inquiry*, *25*(2), 147–186. DOI: https://doi.org/10.1080/1047840X.2014.877340

McElroy, S. E., Rice, K. G., Davis, D. E., Hook, J. N., Hill, P. C., Worthington, E. L., & **Van Tongeren, D. R.** (2014). Intellectual Humility: Scale Development and Theoretical Elaborations in the Context of Religious Leadership. *Journal of Psychology and Theology*, *42*(1), 19–30. Spring. DOI: https://doi.org/10.1177/009164711404200103

Moskowitz, G. B., & **Li, P.** (2011). Egalitarian goals trigger stereotype inhibition: A proactive form of stereotype control. *Journal of Experimental Social Psychology*, *47*(1), 103–116. DOI: https://doi.org/10.1016/j.jesp.2010.08.014

Moss-Racusin, C. A., van der Toorn, J., Dovidio, J., Brescoll, V. L., Graham, M. J., & **Handelsman, J.** (2014). Scientific Diversity Interventions. *Science*, *343*, 615–616. DOI: https://doi.org/10.1126/science.1245936

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ratliff (Ranganath), K. A., Smith,

C. T., Olson, K. R., Chugh, D., Greenwald, A. G., & **Banaji, M. R.** (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*, *18*, 36–88. DOI: https://doi.org/10.1080/10463280701489053

Oswald, F. L., Mitchell, 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. DOI: https://doi.org/10.1037/a0032734

Peirce, J. W. (2007). PsychoPy – Psychophysics software in Python. *J Neurosci Methods*, *162*(1–2), 8–13. DOI: https://doi.org/10.1016/j.jneumeth.2006.11.017

Plant, E. A., & **Peruche, B. M.** (2005). The consequences of race for police officers' responses to criminal suspects. *Psychological Science*, *16*, 180–183. DOI: https://doi.org/10.1111/j.0956-7976.2005.00800.x

Saul, J. (2013). Implicit Bias, Stereotype Threat and Women in Philosophy in Women in Philosophy: What Needs to Change? Edited by Fiona Jenkins and Katrina Hutchison, Oxford University Press, 39–60. DOI: https://doi.org/10.1093/acprof:oso/9780199325603.003.0003

Schlachter, S., & **Rolf, S.** (2017). Using the IAT: how do individuals respond to their results? *International Journal of Social Research Methodology*, *20*(1), 77–92. DOI: https://doi.org/10.1080/13645579.2015.1117799

Schwlenker, J. (2017). What can we learn from the Implicit Association Test? *A Brains Blog Roundtable*, http://philosophyofbrains.com/2017/01/17/how-can-we-measure-implicit-bias-a-brains-blog-roundtable.aspx accessed 04/02/2019

Stafford, T., & **Scaife, R.** OpenIAT. https://osf.io/i8yj5/, DOI: https://doi.org/10.17605/OSF.IO/I8YJ5

Strawson, P. F. (1962). Freedom and Resentment. *Proceedings of the British Academy*, *48*, 1–25. Reprinted in Fischer and Ravizza, 1993. DOI: https://doi.org/10.7591/9781501721564-002

Uhlmann, E. L., & **Cohen, G.** (2007). I Think, Therefore It's True: Effects of Self-Perceived Objectivity on Hiring Discrimination. *Organizational Behavior and Decision Processes*, *104*, 207–223. DOI: https://doi.org/10.1016/j.obhdp.2007.07.001

Vargas, M. (2017). Implicit Bias, Responsibility and Moral Ecology. In D. Shoemaker (ed.), *Oxford Studies in Agency and Responsibility*. DOI: https://doi.org/10.1093/oso/9780198805601.003.0012

Wallace, R. J. (1994). Responsibility and the moral sentiments. Harvard University Press. DOI: https://doi.org/10.2307/2956371

Washington, N., & **Kelly, D.** (2016). Who is responsible for this? In M. Brownstein & J. Saul (Eds.), *Implicit Bias and Philosophy* (pp. 11–36). Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780198766179.003.0002

Webb, T. L., Sheeran, P., & **Pepper, J.** (2010). Gaining control over responses to implicit attitude tests: Implementation intentions engender fast responses on attitude-incongruent trials. *British Journal of Social Psychology*, 1–20. DOI: https://doi.org/10.1348/014466610X532192

Yen, J., Durrheim, K., & **Tafarodi, R. W.** (2018). 'I'm happy to own my implicit biases': Public encounters with the implicit association test. *British Journal of Social Psychology*, *57*(3), 505–523. DOI: https://doi.org/10.1111/bjso.12245

Peer review comments
The author(s) of this paper chose the Open Review option, and the peer review comments can be downloaded at:
http://doi.org/10.1525/collabra.251.pr

How to cite this article: Scaife, R., Stafford, T., Bunge, A., & Holroyd, J. (2020). To Blame? The Effects of Moralized Feedback on Implicit Racial Bias. *Collabra: Psychology,* 6(1): 30. DOI: https://doi.org/10.1525/collabra.251

Senior Editor: Simine Vazire

Editor: Simine Vazire

Submitted: 04 April 2019 **Accepted:** 11 June 2020 **Published:** 03 July 2020

UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Collabra: Psychology is a peer-reviewed open access journal published by University of California Press.

OPEN ACCESS