Journal of Experimental Social Psychology 84 (2019) 103812

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

Consequences of attributing discrimination to implicit vs. explicit bias☆

Natalie M. Daumeyer*, Ivuoma N. Onyeador, Xanni Brown, Jennifer A. Richeson

Yale University, Department of Psychology, United States of America

ARTICLE INFO

Keywords:
Implicit bias
Bias attribution
Accountability
Science communication

ABSTRACT

Implicit bias has garnered considerable public attention, with a number of behaviors (e.g., police shootings) attributed to it. Here, we present the results of 4 studies and an internal meta-analysis that examine how people reason about discrimination based on whether it was attributed to the implicit or explicit attitudes of the perpetrators. Participants' perceptions of perpetrator accountability, support for punishment, level of concern about the bias, and support for various efforts to reduce it (e.g., education) were assessed. Taken together, the results suggest that perpetrators of discrimination are held less accountable and often seen as less worthy of punishment when their behavior is attributed to implicit rather than to explicit bias. Moreover, at least under some circumstances, people express less concern about, and are less likely to support efforts to combat, implicit compared with explicit bias. Implications for efforts to communicate the science of implicit bias without undermining accountability for the discrimination it engenders are discussed.

Since 2015, implicit bias has been mentioned in over 13,000 news articles, currently yields over 40 million Google search results, and was even featured in a recent episode of the popular television show *Grey's Anatomy* (Clack, Rhimes, & Sullivan, 2018). Despite this public attention, little is known about how people incorporate information about implicit bias into their reasoning about discrimination. While many researchers, activists, and reporters presume that educating the public about implicit bias will galvanize support to combat its discriminatory consequences, there is reason to expect greater awareness of implicit bias may *reduce* the extent to which people hold others accountable for the discrimination it engenders (Cameron, Payne, & Knobe, 2010). The present research considers this question.

1. Implicit vs. explicit bias and moral reasoning

Implicit biases are associations and reactions that emerge automatically and often without awareness upon encountering a relevant stimulus (Gawronski & Bodenhausen, 2006; Greenwald & Banaji, 1995). In the social domain, for instance, stereotypical concepts such as "criminal" and "dangerous" automatically come to mind for many people when they encounter or are exposed to images of young black men (Eberhardt, Goff, Purdie, & Davies, 2004; Hester & Gray, 2018). And, people typically unconsciously hold more negative attitudes or feelings about racial/ethnic outgroup, compared with ingroup, members (Axt, Ebersole, & Nosek, 2014). These implicit forms of bias stand in contrast to more explicit forms (Carter & Murphy, 2015; Sommers & Norton, 2006), including preferences, beliefs, and attitudes of which people are generally consciously aware and can, when willing, identify and communicate to others (Dovidio & Gaertner, 2010). Importantly, research suggests that both implicit and explicit bias can shape the judgments and decisions we make (e.g., Bertrand & Mullainathan, 2004; Devine, 1989), as well as how we behave, albeit often in different ways (Dasgupta, 2004; Dovidio, Kawakami, & Gaertner, 2002). But, whereas people are generally aware of the influence their explicit attitudes have on their behavior, they are often unaware of the influence their implicit biases can have (Dovidio et al., 2002).

Conscious awareness, in other words, is a key element that distinguishes implicit from explicit bias. In turn, it is possible that people are likely to hold others less accountable for discriminatory behavior that is thought to be due to implicit, rather than explicit, attitudes. Theoretical and empirical work on moral reasoning argues that in order to be morally responsible for an action, the actor needs to have some level of awareness of and control over their behavior (e.g., Alicke, 2000; Nadler & McDonnell, 2012; Shaver, 1985). Similarly, both classic and more recent models of behavioral attribution (Heider, 1958; Malle, 1999; Malle, Guglielmo, & Monroe, 2014; Monroe & Malle, 2017; Weiner, 1995) assert the critical role of perceived intent in shaping these judgements. In other words, the mental state of the actor is essential to an assessment of culpability (Cameron et al., 2010; Cushman, 2015). Consequently, we generally judge perpetrators of harmful actions more

☆ This paper has been recommended for acceptance by Sarah J Gervais.

* Corresponding author.
E-mail address: natalie.daumeyer@yale.edu (N.M. Daumeyer).

https://doi.org/10.1016/j.jesp.2019.04.010
Received 2 August 2018; Received in revised form 28 March 2019; Accepted 29 April 2019
Available online 24 July 2019

negatively when they are thought to have engaged in those acts knowingly, consciously, and/or intentionally (see Heider, 1958; Knobe & Nichols, 2011).

How might implicit bias attributions for discrimination shape judgments of accountability because of inferences about the mental state of the perpetrator? To the extent that implicit bias is understood as being largely unconscious (i.e., outside of awareness) and/or gives rise to behavior that is unintentional, these theories of attribution suggest that people who engage in discrimination that is due to implicit bias should be held less culpable for their actions than those who engage in discrimination due to explicit bias. Consistent with this prediction, research has found an actor's awareness of their bias to shape the extent to which perceivers hold them culpable for discrimination born of that bias (Cameron et al., 2010; Redford & Ratliff, 2016). When presented with scenarios wherein a manager is said to discriminate against Black people when making hiring or promotion decisions, for instance, perceivers found the manager less morally responsible (blameworthy, accountable) if he was said to be unaware, rather than aware, of his negative attitudes toward Black people (Cameron et al., 2010; Redford & Ratliff, 2016).

Building on this work, the present research considers how people respond to communications of scientific findings that reveal the discriminatory effects of implicit (rather than explicit) bias. Similar to these recent studies, as well as how researchers often discuss implicit bias in the literature (Casad, Flores, & Didway, 2013; Greenwald & Banaji, 1995) and to the public (Payne, Niemi, & Doris, 2017), we focus on awareness (or the lack thereof) as the key factor that distinguishes implicit from explicit forms of bias. Rather than asking participants to respond to a scenario describing discrimination by a single perpetrator (see, again, Cameron et al., 2010; Redford & Ratliff, 2016), however, we present participants with an ostensible news article that details the findings of recent research wherein patterns of discriminatory behavior were observed by categories of perpetrators (e.g., doctors, police officers) and attributed either to their implicit or explicit attitudes. To the extent that implicit bias is understood to involve a lack of awareness of one's attitudes (and/or its influence on behavior), even media reports of scientific studies revealing systemic discrimination should result in the perpetrators being held less accountable if their acts are attributed to implicit, rather than explicit, bias.

The present research also sought to extend the relatively nascent body of research on the consequences of implicit, compared with explicit, bias attribution in two additional ways. First, instead of the overt cases of racial discrimination (e.g., failure to hire/promote Black people) primarily considered in past work, the present research examined responses to subtle differential behavioral treatment in the context of professional (i.e., medical, law enforcement) interactions. Second, in addition to examining the perceived responsibility of perpetrators (i.e., accountability, support for punishment), we also explored whether perceivers report differential concern about implicit compared to explicit bias, support reform efforts to mitigate the effects of implicit compared to explicit bias on behavior, and hold relevant institutional actors (e.g., police departments) differentially accountable for discrimination that is attributed to implicit, relative to explicit, bias. Perhaps, while implicit bias may reduce *individual* culpability for discriminatory behavior, it may not affect the perceived responsibility of *institutional* actors to combat discrimination that occurs in their institutions.

2. The present work

Four studies and an internal meta-analysis consider how people reason about communication of the science of implicit (vs. explicit) bias across a variety of social identity dimensions (political affiliation, age, race) and settings (interactions between doctors and patients, police officers and citizens). Consistent with past work (Cameron et al., 2010; Redford & Ratliff, 2016), we predicted that perpetrators of discrimination would be held less accountable and less worthy of punishment when their discrimination is attributed to their implicit rather than explicit attitudes. We also considered whether this bias attribution shapes perceivers' level of concern about discrimination and the extent to which they support reform efforts to combat it.

3. Studies 1A and 1B

Studies 1A and 1B examined the effects of implicit vs. explicit bias attribution in the context of discrimination by medical doctors toward their patients based on political attitudes. For each study, data collection was completed prior to analyzing the data.

3.1. Methods

All materials and data for all studies are available at OSF (osf.io/5gtyv). All measures, manipulations, and exclusions in the studies are reported in the manuscript.

3.1.1. Participants

Through TurkPrime (Litman, Robinson, & Abberbock, 2017), we recruited Amazon Mechanical Turk workers in the United States to complete each study in exchange for $1.00. In Study 1A, 273 were recruited, however two participants were excluded for admitting to answering randomly and four were excluded due to suspicion, resulting in a sample of 267 participants (43.8% female, 71.5% White, M_{age} = 34.92; 151 in the explicit condition). In Study 1B, 300 participants were recruited, but one was excluded for admitting to answering randomly, leaving in a sample of 299 participants (47.5% female, 74.2% White, M_{age} = 35.00; 149 in the explicit condition). The sample size in Study 1A was chosen based on an a priori decision to enroll a sufficient number of participants to yield an analysis sample with at least 100 people per condition. This target number was increased to 150 for Study 1B, given that it was a replication attempt.

3.1.2. Bias manipulation

Participants read an article inspired by recent research about doctors who demonstrate bias toward their patients based on political ideology (Hersh & Goldenberg, 2016). The article described a study in which both Democratic and Republican doctors exhibited bias toward patients who engaged in somewhat politicized health behaviors—gun ownership and recreational marijuana use. In both conditions, the behavioral consequences of the bias were identical: doctors spent less time with and exhibited more aggressive body language toward patients they had biases against (i.e., marijuana users or gun owners). Similar to past work, the awareness dimension of explicit and implicit bias was manipulated (Cameron et al., 2010; Redford & Ratliff, 2016). In the explicit bias condition, participants read that “doctors were somewhat aware they were treating patients differently, but thought they needed to be tough with their patients in order to encourage them to re-evaluate their behavioral choices and ultimately live a healthier lifestyle.” In the implicit condition, implicit bias was first defined as “attitudes or stereotypes that affect our understanding, actions, and decisions in ways that we are typically not aware of.” Participants also read that “the doctors had no conscious knowledge that they were treating patients any differently based on their political views.”

3.1.3. Dependent variables

All items were rated on Likert scales from 1 (strongly disagree) to 7 (strongly agree).

3.1.3.1. Accountability. Accountability for bias was measured using six items (α_{1A} = 0.83; α_{1B} = 0.82; e.g., “Doctors should be held responsible for any biases they have that may impact how they interact with patients”). Higher scores indicate greater accountability.[1]

3.1.3.2. Punishment. Punishment for bias was measured using three items ($\alpha_{1A} = 0.86$; $\alpha_{1B} = 0.89$; e.g., "Doctors who repeatedly demonstrate biases toward patients should have their license to practice medicine suspended"). Higher scores indicate greater support for punishment.

3.1.3.3. Concern. Concern about the bias was measured using six items ($\alpha_{1A} = 0.84$; $\alpha_{1B} = 0.86$; e.g., "The bias I read about in the article is concerning"). Higher scores indicate greater concern.

3.1.3.4. Reform. Support for reform was measured using five items ($\alpha_{1A} = 0.82$; $\alpha_{1B} = 0.83$; e.g., "Doctors should be required to undergo training to prevent their biases from impacting their treatment of patients"). Higher scores indicate greater support for reform.

3.1.4. Potential moderator variables

We administered several potential moderator variables (e.g., bias awareness, Perry, Murphy, & Dovidio, 2015). Across all studies, none of these scales consistently moderated the effects reported below, and thus, they are not discussed further. Again, for all materials and data see osf.io/5gtyv.

3.2. Procedure

Participants provided informed consent, and then were randomly assigned to the explicit or implicit bias attribution condition. After reading the article, they reported their perceptions of accountability, concern about the bias, support for reform, and support for punishment, in this order. Items within scales were randomized across participants. Participants next completed a number of potential moderator variables, then reported their demographic information (e.g., race, age, gender, political ideology, and whether they own a gun or engage in recreational marijuana use). Participants also completed data quality and attention checks, and then were thanked, debriefed, and paid.

3.3. Results

We conducted an independent samples *t*-test exploring the effect of the bias attribution manipulation (explicit vs. implicit) on each dependent measure.[2] For each outcome, we present the results for Study 1A followed by those for Study 1B (the replication sample). For Study 1A, the analyses had 80% power to detect an effect size of Cohen's $d = 0.35$. For Study 1B, the analyses had 80% power to detect an effect size of $d = 0.33$. Correlations among the dependent variables for each study are provided in Table 1. The means and standard deviations for each dependent variable by condition are provided in Table 2.

3.3.1. Accountability

Analyses revealed that there was unequal variance for perceptions of accountability among the two conditions in Study 1A, $F(1,265) = 6.46$, $p = .012$. Thus, we provide *t*- and *p*-values for unequal variances assumed. Consistent with predictions, there was a significant difference in accountability as a function of bias attribution in Study 1A, $t(221.39) = 3.51$, $p = .001$, $d = 0.44$. Participants in the implicit bias condition held doctors less accountable than participants in the explicit bias condition. This effect replicated in Study 1B, $t(297) = 5.00$, $p < .001$, $d = 0.58$.

[1] In Study 1A, a scale item referred to "implicit" bias, however, no participant noticed nor do the results change with its exclusion from the composite. This was corrected in Study 1B.

[2] Analyses in all studies are robust to the inclusion of participant gender, age, race, and conservatism as covariates.

Table 1
Correlations of dependent variables for studies 1A, 1B, and 2.

	Study 1A			Study 1B			Study 2		
	1.	2.	3.	1.	2.	3.	1.	2.	3.
1. Accountability									
2. Punishment	0.57			0.56			0.52		
3. Concern	0.70	0.64		0.69	0.58		0.72	0.54	
4. Reform	0.59	0.66	0.58	0.57	0.62	0.55	0.63	0.58	0.63

All correlations are significant at the $p < .01$ level.

3.3.2. Punishment

Contrary to predictions, in Study 1A, there was no difference in support for punishment as a function of bias attribution, $t(265) = 1.01$, $p = .312$, $d = 0.13$. In the replication sample (Study 1B), however, support for punishment did differ by condition, $t(297) = 3.91$, $p < .001$, $d = 0.45$. Participants in the implicit bias condition were significantly less supportive of punishment than participants in the explicit bias condition.

3.3.3. Concern

In Study 1A, there was no difference in participants' concern about bias as a function of condition, $t(265) = 1.08$, $p = .281$, $d = 0.13$. In Study 1B, however, concern about bias did differ significantly by condition, $t(297) = 3.42$, $p = .001$, $d = 0.40$, such that participants in the implicit bias condition expressed less concern than participants in the explicit bias condition.

3.3.4. Reform

In Study 1A, there was no difference in support for reform efforts as a function of bias attribution, $t(265) = 0.80$, $p = .427$, $d = 0.10$. The effect in Study 1B was also not reliable, $t(297) = 1.43$, $p = .155$, $d = 0.16$.

3.4. Discussion

Studies 1A and 1B provide initial evidence that attributing systemic discrimination to implicit rather than explicit bias results in lower perceived perpetrator accountability, consistent with past research (Cameron et al., 2010). Study 1B also revealed lower support for punishment and concern about the bias when discrimination is attributed to implicit rather than explicit beliefs. Interestingly, support for reform efforts to combat the influence of bias in medical interactions did not differ among participants in the implicit and explicit bias conditions.

4. Study 2

Studies 1A and 1B provide evidence that attributing discrimination to implicit versus explicit bias reduces perceptions of perpetrator accountability and, perhaps also, support for punishment and concern about bias and discrimination. The outcomes of the discrimination in these studies, however, were relatively modest—unsatisfactory doctor-patient interactions. Reduced accountability for discrimination that is attributed to implicit bias may not emerge when the outcome is more harmful. Alternatively, even when harm is severe, people excuse seemingly unintentional behaviors (Ames & Fiske, 2013). Thus, implicit bias may still undermine accountability. Study 2 sought to investigate this possibility. Specifically, Study 2 sought to discern whether reduced accountability for implicit, relative to explicit, bias attribution emerges for discriminatory outcomes that are more harmful than the unsatisfactory interactions described in Study 1. Study 2 also examined these questions in a second bias domain; namely, ageism.

Table 2
Means and Standard Deviations of Dependent Variables by Condition for Studies 1A, 1B, and 2.

	Study 1A				Study 1B				Study 2			
	Explicit $N = 151$		Implicit $N = 116$		Explicit $N = 149$		Implicit $N = 150$		Explicit $N = 209$		Implicit $N = 217$	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Accountability	5.35^{a}	0.97	4.88^{b}	1.17	5.39^{a}	0.99	4.81^{b}	1.00	5.50^{a}	0.93	5.15^{b}	1.03
Punishment	4.51^{a}	1.49	4.32^{a}	1.51	4.83^{a}	1.54	4.14^{b}	1.49	4.94^{a}	1.29	4.68^{b}	1.37
Concern	4.72^{a}	1.21	4.56^{a}	1.18	4.95^{a}	1.16	4.50^{b}	1.12	5.16^{a}	1.07	4.93^{b}	1.07
Reform	5.06^{a}	1.14	4.94^{a}	1.20	5.17^{a}	1.19	4.98^{a}	1.15	5.53^{a}	0.96	5.32^{b}	1.10

Means with different superscripts within each study and variable differ significantly at the $p < .05$ level. The means for Study 2 are collapsed across levels of harm.

4.1. Method

4.1.1. Participants

Through Amazon Mechanical Turk, 427 United States residents participated in exchange for $1.00. One person reported suspicion of the manipulation, leaving a total of 426 participants (53.1% female, 68.5% White, $M_{age} = 36.28$, 209 in the explicit condition; 211 in the low harm condition).[3] For Study 2, we again sought to enroll a sufficient number of participants to yield an analysis sample with at least 100 people in each condition.

4.1.2. Bias manipulation

As in the previous two studies, participants read an article about research finding that doctors demonstrated bias in their behavior toward their patients, this time, ageism toward older patients (see, again, osf.io/5gtyv). In the article, ageism was defined as beliefs associating older adults with incompetence. Awareness was again the primary marker of implicit vs. explicit bias. In the explicit condition participants read, "The doctors were aware that they had these negative beliefs, and felt they were valid because they were based on actual experiences and interactions with older adult patients." In the implicit condition participants read, "The doctors had no conscious knowledge of their implicit ageism nor that their ageism was affecting their treatment of patients based on age." In order to ensure participants knew that the doctors were "well-meaning" in both conditions, each article also reported that: "Nearly all of the doctors in the study thought that they treated all of their patients with the same level of care, concern, and attention."

4.1.3. Harm manipulation

In both conditions, the initial outcomes associated with the doctors' behavior were similar to those described in Study 1: "Doctors with more ageist attitudes tended to spend less time with and demonstrate more dismissive body language toward older patients."

In the *low-harm* condition, participants also read, "Furthermore, the researchers found that older patients of doctors with high levels of age bias reported being less satisfied with their medical encounter than patients of doctors with little to no age bias." In addition, the low-harm condition article ended with a statement highlighting that "age bias could be a potential problem for patient care, including leading to undesirable patient interactions."

In the *high-harm* condition participants read, "Furthermore, this [poor treatment] led to older patients getting less medical attention than they needed. Of greater concern, the researchers found that older patients of doctors with high levels of age bias do not live as long as older patients of doctors with little to no age bias." In addition, the high-harm condition article concluded with a statement highlighting that, "age bias could be a potential problem for patient care, including leading to premature death."

[3] 4.0% of participants reported a medical profession (e.g., nurse), however, their exclusion did not alter the results.

4.1.4. Dependent variables

The dependent measures were the same as described in Study 1B, but modified to refer to ageism rather than political bias. In addition, the concern scale was increased to seven items in order to include statements reflecting concern about the bias itself (e.g., "The bias I read about in the article is concerning"), as in our previous studies, and concern about the outcomes of the bias (e.g., "The treatment I read about in the article is concerning"). The items nevertheless formed a reliable single concern scale ($\alpha = 0.87$), with higher scores reflecting greater concern.

4.2. Procedure

After providing informed consent, participants read one of the four articles (i.e., explicit bias/low-harm; explicit bias/high-harm; implicit bias/low-harm; implicit bias/high-harm), based on random assignment. They next completed the primary outcome variables, and provided relevant demographic information (e.g., age, race, gender, occupation, political conservatism). Last, participants answered the data quality and suspicion check questions, were thanked, debriefed, and paid.

4.3. Results

The dependent variables were thus subjected to a 2 (bias type: explicit vs. implicit) × 2 (harm: low vs. high) analysis of covariance. The sample provided 80% power to detect an effect size of $\eta^2_{partial} = 0.019$. Given the bias domain featured in this study, however, analyses exploring moderation by participant age are provided in the supplemental materials. Correlations among the dependent measures are provided in Table 1. The means and standard deviations are provided in Table 2.

4.3.1. Accountability

Consistent with predictions and replicating Studies 1A and 1B, the main effect of bias attribution on accountability was significant, $F(1,422) = 13.66$, $p < .001$, $\eta^2_{partial} = 0.031$. Unexpectedly, the main effect of harm was not, $F(1,422) = 0.55$, $p = .459$, $\eta^2_{partial} = 0.001$; and, importantly, the interaction between bias attribution and harm was also not significant, $F(1,422) = 0.04$, $p = .837$, $\eta^2_{partial} < 0.001$. That is, participants who read about implicit bias held the doctors significantly less accountable than participants who read about explicit bias, regardless of the level of harm associated with the discriminatory behavior.

4.3.2. Punishment

Consistent with Study 1B and as depicted in Fig. 1, participants in the implicit bias condition supported punishing the doctors less than participants in the explicit bias condition, $F(1,422) = 4.26$, $p = .040$, $\eta^2_{partial} = 0.010$. The main effect of harm was also significant, $F(1,422) = 8.48$, $p = .004$, $\eta^2_{partial} = 0.020$, with high harm resulting in greater support for punishment than low harm. Again, however, the interaction between bias attribution and harm was not significant, *F*

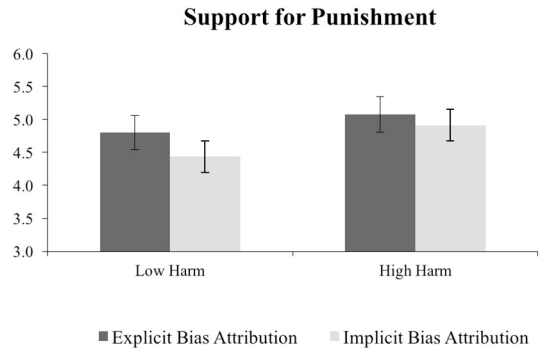

Fig. 1. Bias attribution and harm level on support for punishment. The effect of bias attribution condition (explicit vs. implicit) and harm condition (low vs. high) on support for punishment in Study 2. Error bars represent 95% confidence intervals.

(1,422) = 0.61, p = .435, $\eta^2_{partial}$ = 0.001.

4.3.3. Concern

Replicating Study 1B, participants in the implicit bias condition reported less concern about the bias (and its outcomes) than participants in the explicit bias condition, $F(1,422) = 5.20$, $p = .023$, $\eta^2_{partial} = 0.012$. Neither the main effect of harm, $F(1,422) = 0.23$, $p = .633$, $\eta^2_{partial} = 0.001$, nor the bias attribution by harm interaction, $F(1,422) = 0.61$, $p = .435$, $\eta^2_{partial} = 0.001$, were significant.

4.3.4. Reform

Unlike in the previous studies, support for reform differed significantly as a function of bias attribution, $F(1,422) = 4.73$, $p = .030$, $\eta^2_{partial} = 0.011$. Participants in the implicit bias condition expressed less support for reform than participants in the explicit bias condition. There was no main effect of harm, $F(1,422) = 0.52$, $p = .470$, $\eta^2_{partial} = 0.001$, nor an interaction between bias attribution condition and harm, $F(1,422) = 0.18$, $p = .673$, $\eta^2_{partial} < 0.001$.

4.4. Discussion

Study 2 provides additional evidence that when discrimination is attributed to implicit rather than explicit bias, people hold the perpetrators less accountable. Notably, even doctors whose discrimination was linked to premature patient death (i.e., severe harm) were held less accountable when the doctors were said to be unaware (implicit condition), rather than aware (explicit condition) of their ageist beliefs. Replicating Study 1B, Study 2 participants expressed less support for punishing perpetrators and less concern about the bias in the implicit compared with explicit bias condition. They also revealed differential support for reform efforts when the discrimination described was attributed to implicit, rather than explicit, bias. These findings suggest that the communication of scientific studies detailing the effects of implicit bias on behavior may unwittingly increase tolerance for both discriminators and discrimination itself, even when the harm is quite severe.

5. Study 3

Given our concern about potential unexpected effects of increased public attention on implicit bias, Study 3 sought to extend our examination to a domain in which the role of implicit bias in discriminatory or otherwise disparate outcomes has penetrated public consciousness; namely, police misconduct against racial minority citizens. Specifically, we explored whether attributing racially disparate outcomes of police-citizen encounters to implicit, rather than explicit, bias reduces perceptions of police accountability.

Because of the context of the discrimination—namely, socially sensitive and race-related, we explored whether individual differences in concern about appearing racially prejudiced might affect judgments of the police officers either in general or as a function of the implicit vs. explicit bias attribution. Specifically, participants completed Plant and Devine's (1998) measures of internal (IMS) and external (EMS) motivation to respond without prejudice. IMS assesses the extent to which people are intrinsically motivated to behave in non-prejudiced ways (e.g., to live up to their values), whereas EMS assesses the extent to which they are externally motivated to respond in non-prejudiced ways (e.g., to avoid public condemnation). Given that participants made their responses to the scenarios in the present work anonymously (i.e., online), we did not expect EMS to be relevant to these judgments.

How might IMS moderate judgments of perpetrators who discriminate because of implicit, compared to explicit, racial bias? Past research has found that IMS can shape reactions to one's own displays of racial bias, most notably feelings of compunction for failing to regulate one's implicit bias successfully (Devine, Monteith, Zuwerink, & Elliot, 1991; Devine, Plant, Amodio, Harmon-Jones, & Vance, 2002; Legault, Green-Demers, Grant, & Chung, 2007; Plant & Devine, 1998). When confronted with the possibility that one harbors implicit bias, for instance, people with high, relative to low, levels of IMS attempt to reduce their implicit bias through training (Plant & Devine, 2009) and subsequently behave in ways that are less discriminatory (Cooley, Lei, & Ellerkamp, 2018). Based on this work, it is possible that individuals higher (compared with lower) in IMS may not show reduced accountability for people who discriminate due to implicit bias relative to those who discriminate due to explicit bias.

A quite different pattern of results, however, is also possible. Specifically, one hallmark of being internally motived to respond without prejudice is harboring low levels of explicit racial bias (Plant & Devine, 1998), and certainly doing one's best not to behave in discriminatory ways due to any explicitly-held biases (Legault et al., 2007; Plant & Devine, 2009). Coupled with general societal condemnation of explicit forms of racial bias (Crandall & Eshleman, 2003), it is possible that individuals with higher levels of IMS may be especially likely to hold perpetrators of discrimination that is thought to be due to explicit bias *more* accountable and see them as more blameworthy than perpetrators of implicit bias, whereas individuals lower in IMS may be less likely to distinguish between individuals who discriminate because of implicit, compared with explicit, racial bias. Such a pattern would suggest, further, that individuals with higher levels of IMS are more sympathetic to perpetrators of racial discrimination that is attributed to implicit, rather than to explicit, racial bias, perhaps because they can readily identify with such individuals. The present work examined these possibilities and, in so doing, is a rare exploration of the extent to which perceivers' own prejudice-related motivations shape moral judgments regarding perpetrators of discrimination.

In addition, Study 3 sought to address ambiguity in the perceived accountability measure. Unlike our previous studies, we assessed perceived accountability for having implicitly (compared to explicitly) biased *beliefs* separately from accountability for biased *behavior*. Biased beliefs are typically thought to be less controllable and, likely, less changeable than behavior (Bargh, 1999; Devine, 1989), and largely outside the domain of moral or legal judgment (Mitchell & Tetlock, 2006). Consequently, perpetrators may assess accountability for implicit vs. explicit attitudes somewhat differently than accountability for the behaviors they engender. In Study 3, we also offered participants an opportunity to differentiate between the culpability they assign the perpetrators themselves and relevant institutional actors. Specifically, we assessed support for punishment and for reform efforts at both the individual (i.e., police officer) and institutional (i.e., police department) levels. It is possible that people hold individual police officers less accountable for discrimination born of implicit, compared with explicit, bias, but hold police departments equally accountable or, even, hold departments even more responsible for combatting their officers' implicit compared with explicit bias.

5.1. Methods

5.1.1. Participants

Through Amazon Mechanical Turk, 227 self-identified White Americans participated in exchange for $1.00. Four participants were excluded due to suspicion, resulting in a sample of 223 White participants (48.4% female, M_{age} = 36.81, 111 in explicit condition). Consistent with our previous studies, we sought to enroll a sample with at least 100 people in each condition.

5.1.2. Bias manipulation

As in the previous studies, participants read an ostensible news article about a scientific study of a large metropolitan area in which police officers were found to behave in racially biased ways during interactions with racial minority citizens. The consequences of the bias were the same regardless of the bias attribution: police officers with greater racial bias were said to behave more aggressively toward and were more likely to handcuff or detain racial minorities compared to Whites. We again manipulated explicit vs. implicit bias attribution by emphasizing the unconscious nature of implicit bias.

5.1.3. Dependent measures

All items were rated on Likert scales from 1 (strongly disagree) to 7 (strongly agree).

5.1.3.1. Accountability. Accountability for having biased *beliefs* was measured with three items (α = 0.76; e.g., "Police officers should be held accountable for their biased racial attitudes and beliefs"), as was accountability for biased *behaviors* (α = 0.73; e.g., "Police officers should be held accountable for how they treat racial minority citizens").

5.1.3.2. Punishment. Support for individual punishment (α = 0.88) was measured by having participants indicate their agreement with five punishments following the stem, "Police officers who consistently demonstrate more negative behavior toward minority, compared with white, citizens should be…" (taken off patrol, fined, suspended without pay, demoted, and fired). Support for institutional-level punishment (α = 0.83) was measured by having participants indicate their agreement with four punishments following the stem, "Police departments with officers who consistently demonstrate more negative behavior toward minority, compared with white, citizens should be…" (investigated, fined, forced to hire a new chief, and taken over by the Justice Department).

5.1.3.3. Reform. Individual reform was measured with three items (α = 0.85; e.g., "Police officers should be required to undergo training to prevent any biased attitudes they may have from impacting their treatment of citizens"). Institutional-level reform was measured with three items (α = 0.79, e.g., "Police departments should work on developing better relationships with racial minority communities to reduce the likelihood of negative officer–citizen interactions").

5.1.3.4. IMS. Internal motivation to respond without prejudice (IMS; Plant & Devine, 1998) was assessed with five items (α = 0.91, e.g., "I am personally motivated by my beliefs to be nonbiased") rated on 1–7 agreement scales. Because IMS was measured across all studies, we adapted the scale to measure motivation to respond without prejudice in general, rather than specifically regarding Black Americans. Higher scores reflect greater levels of internal motivation.

5.2. Procedure

After providing informed consent, participants read one of the randomly assigned bias attribution articles. They then completed the outcome measures, followed by the IMS scale (embedded among a number of other measures). Participants next reported their demographic information (e.g., age, gender, race, conservatism, occupation), followed by data quality, attention, and suspicion checks. After, they were thanked, debriefed, and paid.

Table 3
Correlations of variables for study 3.

	Study 3				
	1.	2.	3.	4.	5.
1. Accountability					
2. Individual Punishment	0.64				
3. Departmental Punishment	0.65	0.79			
4. Individual Reform	0.70	0.53	0.59		
5. Departmental Reform	0.66	0.43	0.53	0.79	
6. IMS	0.59	0.42	0.46	0.59	0.53

All correlations were significant at the $p < .01$ level.

5.3. Results

Preliminary analyses confirmed that IMS scores did not vary systematically due to the bias attribution manipulation, $t(221) = -1.54$, $p = .124$. To test our predictions, then, we subjected each of the dependent measures to a regression with bias attribution condition (explicit = −0.5, implicit = 0.5), IMS scores (centered), and the interaction of these two variables as predictors. The analyses had 80% power to detect an effect size of $R^2_{partial} = 0.028$. Means and standard deviations presented in the text are adjusted for the effects of IMS and the condition × IMS interaction. The correlations among the variables and are provided in Table 3.

5.3.1. Accountability

Because the two measures of accountability (for biased beliefs, for biased behavior) were highly correlated ($r = 0.86$), we averaged the items to form a single accountability composite. Analyses revealed a significant effect of IMS, $B = 0.47$, $SE = 0.04$, $t(219) = 10.92$, $p < .001$, with participants higher in IMS holding the police officers more accountable than participants lower in IMS. The main effect of bias attribution manipulation was just at the threshold of significance, $B = -0.21$, $SE = 0.11$, $t(219) = -1.97$, $p = .050$, and the means were in the predicted direction. Participants who read that the police officers' discrimination was due to implicit bias ($M = 5.44$, $SD = 0.81$) held them less accountable than participants who read that the police officers' discrimination was due to their explicit bias ($M = 5.65$, $SD = 0.81$). This effect was not moderated by IMS, $B = -0.08$, $SE = 0.09$, $t(219) = -1.01$, $p = .314$.

5.3.2. Punishment

Analyses of support for *individual-level* punishment, again, revealed a significant effect of IMS, $B = 0.46$, $SE = 0.07$, $t(219) = 6.84$, $p < .001$; higher levels of IMS predicted greater support for punishing the police officers. Although the means were in the predicted direction, the main effect of bias attribution was not statistically reliable, $B = -0.29$, $SE = 0.17$, $t(219) = -1.75$, $p = .082$, nor was the bias attribution by IMS interaction, $B = -0.13$, $SE = 0.13$, $t(219) = -0.98$, $p = .327$. Participants who read that police officer's discrimination was due to implicit bias ($M = 4.87$, $SD = 1.26$) did not significantly differ in their support for punishing individual policers from participants who read their discrimination was due to explicit bias ($M = 5.17$, $SD = 1.26$).

Support for *institutional-level* punishment (i.e., police departments) was subjected to the same analysis. Once again, the main effect of IMS emerged, with those higher in IMS again expressing greater support for punishing police departments, $B = 0.46$, $SE = 0.06$, $t(219) = 7.48$, $p < .001$. Contrary to predictions, the main effect of bias attribution

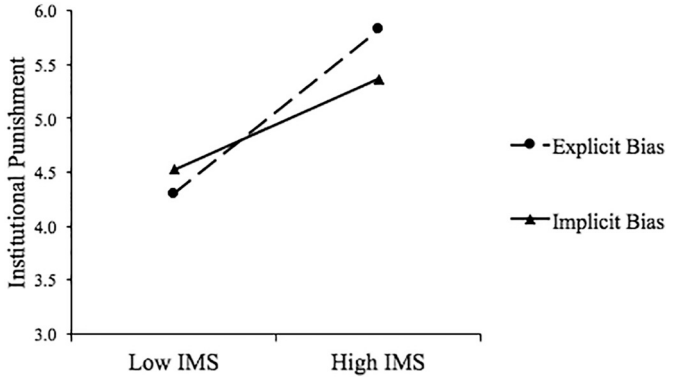

Fig. 2. Support for institutional punishment by bias attribution condition and internal motivation to respond without prejudice (IMS). High and low IMS represent +/− 1 SD from the mean, respectively.

condition was not reliable, $B = -0.12$, $SE = 0.16$, $t(219) = -0.75$, $p = .453$, however, the IMS by condition interaction did reach conventional levels of statistical significance, $B = -0.27$, $SE = 0.12$, $t(219) = -2.15$, $p = .033$.

To decompose this interaction, we used PROCESS (model 1) to test for the effect of bias attribution condition at low (−1 SD from the mean) and high (+1 SD from the mean) levels of internal motivation to respond without prejudice (IMS). Analyses revealed that among those low in IMS, the effect of bias attribution condition was not significant, $B = 0.22$, $SE = 0.22$, $t(219) = 0.99$, $p = .322$. Participants with higher levels of IMS, however, did differentiate between department-level punishment after reading about officer discrimination due to explicit compared with implicit bias, $B = -0.46$, $SE = 0.22$, $t(219) = -2.07$, $p = .039$; as depicted in Fig. 2, high IMS participants expressed greater support for punishing police departments when they read about officers discriminating due to explicit rather than implicit bias. For the full model see supplemental materials.

5.3.3. Reform

Analyses of participants' support for *individual-level* reform efforts for reducing police officer discriminatory behavior revealed a pattern similar to that found for departmental punishment. Specifically, although the effect of bias attribution was not significant, $B = -0.13$, $SE = 0.12$, $t(219) = -1.02$, $p = .309$, a significant main effect of IMS emerged, $B = 0.52$, $SE = 0.49$, $t(219) = 10.55$, $p < .001$, which was qualified by a significant interaction with bias attribution condition, $B = -0.22$, $SE = 0.10$, $t(219) = -2.19$, $p = .030$. We again used PROCESS (model 1) to decompose this interaction, examining the effect of bias attribution at low (−1 SD) and high (+1SD) levels of IMS. Analyses again revealed that among participants low in IMS, the effect of bias condition was not significant, $B = 0.15$, $SE = 0.18$, $t(219) = 0.83$, $p = .405$. However, among participants high in IMS, the effect of bias attribution condition was significant, $B = -0.40$, $SE = 0.178$ $t(219) = -2.29$, $p = .023$. Participants high in internal motivation to respond without prejudice supported reform efforts targeted at individual officers more when discrimination was attributed to explicit, rather than implicit, bias. Again, for the full model see supplemental materials.

Analyses of support for *institutional-level* reform efforts targeted at police departments revealed only a main effect of IMS, $B = 0.40$, $SE = 0.04$, $t(219) = 9.02$, $p < .001$. Regardless of bias attribution condition, participants higher in IMS were more supportive of department-level reform efforts than participants lower in IMS. Neither the bias attribution manipulation [$B = -0.09$, $SE = 0.11$, $t(219) = -0.84$, $p = .401$] nor its interaction with IMS [$B = -0.06$, $SE = 0.09$, $t(219) = -0.72$, $p = .471$] were reliable. Participants who read that police officers discriminate because of implicit bias ($M = 5.69$, $SD = 1.18$) did not significantly differ from participants who read that police officers discriminate because of explicit bias ($M = 5.43$, $SD = 1.37$) in their support for reform efforts aimed at police departments.

5.4. Discussion

In a domain that has received considerable public attention—racially disparate patterns of police misconduct—Study 3 revealed that attributing discrimination to implicit rather than explicit bias reduces perceived perpetrator accountability. Further, individual differences in internal motivation to respond without prejudice (IMS) predicted overall assessments of perpetrator accountability, support for punishment, and support for reform. Specifically, high IMS participants tended to hold the police officers more accountable for discriminatory behavior than low IMS participants, irrespective of the bias attribution. For institutional-level punishment and individual-level reform, interestingly it was the high IMS participants who differentiated between discrimination based in implicit and explicit racial bias. That is, participants with higher levels of IMS supported more severely punishing police departments and more reform efforts for individual officers when police officer bias was said to be explicit, rather than, implicit. Participants lower in IMS did not reveal these differences. Taken together, these findings suggest that judgments beyond perpetrator accountability for discrimination due to implicit vs. explicit bias may be shaped, at least in part, by domain-relevant motivations of perceivers. Future research is, of course, necessary to replicate and explore this finding more deeply.

6. Internal meta-analysis

In order to assess the robustness of the effects observed across our studies, we conducted an internal meta-analysis (Goh, Hall, & Rosenthal, 2016). Specifically, we calculated Cohen's *d* for the explicit vs. implicit bias attribution effect using the means and standard deviations for each dependent variable in each study. For Study 2, we collapsed across harm to generate the estimates. For Study 3, we created composites for punishment and reform that collapsed across the individual- and institutional-level distinction, and then re-calculated the bias attribution manipulation effect-size estimates for each. Given there were no significant main effects of bias attribution on punishment and reform in Study 3, this decision should provide conservative estimates of each effect-size. We also used the adjusted means and standard deviations (controlling for IMS and the condition × IMS interaction) to calculate Cohen's *d*. Analyses revealed a significant effect of accountability ($d = 0.41$, $Z = 7.08$, $p < .001$). Across all studies, when participants read that discriminatory behavior was due to implicit compared to explicit bias, they perceived less perpetrator accountability (see Fig. 3). Additionally, analyses revealed modest, albeit statistically significant, overall effect-size estimates for each of the other outcome

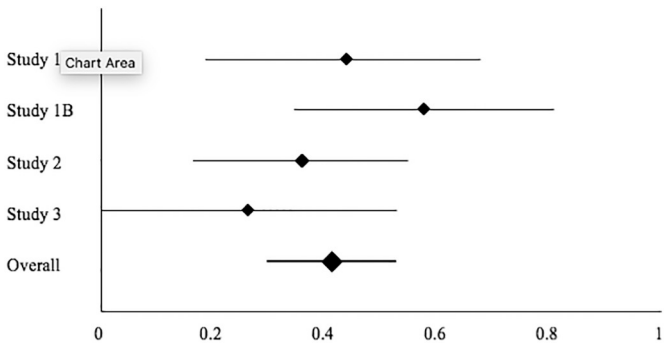

Fig. 3. Effect sizes for accountability across studies. Cohen's *d* for accountability plotted for each study. Error bars represent 95% confidence intervals.

variables: punishment ($d = 0.24$, $Z = 4.20$, $p < .001$), concern ($d = 0.25$, $Z = 3.90$, $p < .001$), and reform ($d = 0.16$, $Z = 2.81$, $p = .005$).

7. General discussion

Scholars of stereotyping and prejudice have long expressed concern that emphasizing the role of unconscious, automatic beliefs in engendering discrimination might reduce the perceived culpability of its perpetrators (Fiske, 2004). The present findings suggest this concern was well-founded. Specifically, in the context of scientific research communications, we found that people hold perpetrators less accountable for discriminatory behavior when it is attributed to their implicit, rather than explicit, attitudes. This reduced accountability effect was observed in two contexts (medical and police interactions), across three different biases (political, age-based, racial), and, somewhat surprisingly, was not attenuated when the consequences of the discrimination were especially severe (i.e., premature death). Given that a great deal of discrimination is rooted in implicit forms of bias (Greenwald & Pettigrew, 2014), the small, but reliable, reduced accountability effect found here could be highly consequential.

In addition to being held less accountable for discrimination, perpetrators were somewhat less likely to be punished if their behavior was described as stemming from implicit rather than explicit bias. Moreover, perceivers expressed lower levels of concern about, and were less likely to support efforts to mitigate against the effects of, implicit compared with explicit bias. Taken together, these findings suggest that communications of scientific findings regarding the effects of implicit bias may unwittingly reduce the perceived severity of the discrimination it engenders.

7.1. Theoretical implications

The primary finding of the present work—reduced accountability for discriminatory acts born of implicit rather than explicit bias—is consistent with prevailing models of moral reasoning, all of which underscore the essential role of a person's mental state in shaping judgments of culpability for harmful actions (Cushman, 2015; Knobe & Nichols, 2011; Malle et al., 2014; see also Heider, 1958). Most notably, an actor needs to have some level of awareness of and control over their behavior in order to be held morally responsible for it (e.g., Alicke, 2000; Shaver, 1985). Hence, as we observed here, people judge perpetrators of harmful actions more negatively when they are thought to have engaged in those acts knowingly, consciously, and/or intentionally (Monroe & Malle, 2017, 2019; see also Cameron et al., 2010; Redford & Ratliff, 2016). In the present work, that is, perpetrators of discrimination who were said to be influenced by attitudes and beliefs that they held unconsciously (i.e., implicitly) were held less accountable than perpetrators of the same discriminatory behavior who were said to be influenced by attitudes and beliefs that they held consciously (i.e., explicitly).

While our results may not be surprising based on this larger theoretical work, they are particularly compelling given that the behaviors examined across these studies are far more complex than the types of behaviors typically studied in the moral reasoning literature. Specifically, relatively rich descriptions of several behaviors are provided in the articles used in the present studies rather than exposing participants to discrete behaviors that target individuals ostensibly have engaged in. In addition, many of the behaviors described in our work can surely be classified as conscious (or even intentional) acts, including aggressive behavior toward citizens by police officers. That is, on their own, they could and perhaps should be understood as deserving of the level of moral condemnation associated with intentional behavior. Yet, describing the differential behaviors observed as stemming from implicit, rather than explicit, bias seems to offer perceivers an adequate basis for reducing judgments of accountability.

Should the tenets of prevailing models of moral responsibility apply to these types of behaviors? Imagine that a person regularly and repeatedly steals the morning newspaper from one of his neighbors. Would his perceived accountability for this action be reduced if we later find out that he is unaware (vs. aware) of his negative affect toward the neighbors? It certainly seems unlikely. Similarly, the doctors and police officers described in our work knowingly engaged in at least some of the behaviors described in the scenarios (e.g., aggressive body language). What they are said to be unaware of is the extent to which they were behaving differently toward different types of targets based on their social group memberships. That is, the relation between their attitudes regarding and behaviors toward members of a disfavored, relative to favored, social category. The extent to which current models of moral reasoning should be relevant to these behaviors is not clear, but the present findings suggest that perceivers lower their assessments of accountability nonetheless. Future research is needed to discern exactly when, why, and for whom accountability for even clearly intentional behaviors is reduced by locating a possible origin of the behavior to attitudes and beliefs that are held implicitly rather than explicitly (see Monroe & Malle, 2017, 2019).

7.2. Practical implications

At first glance, one potential takeaway from the present work is for researchers to avoid discussing implicit bias with the public. This is certainly not our position. Indeed, we maintain that the penetration of the implicit bias construct into public consciousness has certainly been a much-needed corrective to outdated models of stereotyping, prejudice, and discrimination that require clear evidence of antipathy, animus, and often discriminatory intent in order to classify behavior as biased (Allport, 1954; Simon, Moss, & O'Brien, 2019; Sommers & Norton, 2006; Swim, Scott, Sechrist, Campbell, & Stangor, 2003; see also Washington v. Davis, 1976). Rather than calling for reduced public discussion of the science of implicit bias, we believe it is time for more nuanced public conversations. For example, it may be time to revisit the tendency for researchers and others to describe implicit bias as unconscious and/or uncontrollable. Not only is there evidence that people do have some ability to detect their implicit biases (Hahn, Judd, Hirsh, & Blair, 2014; Monteith, Voils, & Ashburn-Nardo, 2001; Uhlmann & Nosek, 2012), but the present results suggest that implicit bias attributions—that is, the lack of awareness of bias—engender reduced accountability judgments. In addition, public conversations about implicit bias could also include information about the potential for individuals to override the effects of even implicitly-held attitudes and beliefs on their behavior, at least when they have the opportunity and motivation (Dunton & Fazio, 1997; Nosek, Hawkins, & Frazier, 2011). Public conversations should also highlight the potential for policies and decision-making structures (e.g., diverse hiring panels) to combat the influence of implicit bias on individual acts and judgments (e.g., Daumeyer, Rucker, & Richeson, 2017).

Relatedly, it may be possible to capitalize on the popular knowledge of implicit bias in order to help combat it. At some point, awareness that implicit bias is a common pathway toward the reproduction of unequal and unjust societal outcomes based on race, gender, age, and other classifications, should motivate efforts to combat it structurally and institutionally, in addition to any individual-level efforts (see Kelly & Roedder, 2008). In addition, the more widely-known implicit bias becomes, the more people (and relevant institutions) can and should be held accountable for its effects. That is, when people think that actors should have been able to foresee the harmful consequences of even their unintended actions, they hold those actors more accountable (Lagnado & Channon, 2008; Laurent, Nuñez, & Schweitzer, 2016; Monroe & Malle, 2019). Thus, rather than calling for less public discussion of implicit bias, we argue for a more sophisticated conversation on the ways in which implicit bias shapes behavior and multiple ways to combat it (e.g., Payne, Vuletich, & Lundberg, 2017).

7.3. Limitations

One limitation of the present work is its failure to probe whether the results may differ for people who share a meaningful group membership with the victims of the discrimination. There is reason to expect that the targets of the discrimination may not hold perpetrators any less accountable for discrimination based on an implicit, rather than explicit, bias attribution. Not only would such a finding reveal an important moderator of the accountability effect found here, but it would also suggest that motivation may play a role in engendering what has heretofore largely been understood as a purely reasoned discounting of moral responsibility. The patterns of moderation by internal motivation to respond without prejudice that emerged in Study 3 suggest a role for motivation in shaping these judgments, this question is especially ripe for examination in future research.

It is also important to acknowledge that the primary manipulation in the present work distinguished implicit from explicit bias by manipulating awareness. While this decision is grounded in and, thus, consistent with past work (e.g., Redford & Ratliff, 2016), and as noted previously, reflects how researchers tend to discuss implicit bias both in the literature (Greenwald & Banaji, 1995) and with the public (Payne, Niemi, & Doris, 2018), it is only one of the characteristics that differentiates implicit from explicit bias. Had we focused on other relevant distinguishing characteristics (e.g., controllability, efficiency, automaticity, etc.), it is possible that a different pattern of results may have emerged. For example, "automatic" bias that is described as potentially controllable does not yield reduced moral responsibility judgments relative to explicit bias (Cameron et al., 2010).

Further, discrimination born of implicit bias is sometimes thought to be unintentional or at least less intentional than discrimination born of explicit bias (Cameron et al., 2010; Onyeador, 2017). Thus, had we defined implicit bias by a lack of intentionality, rather than awareness, we would expect similar results to the ones found here. Indeed, they may have been even more robust. Clarifying the role of perceived intent as a potential mediator between holding implicit forms of bias and reduced accountability for behaviors that are said to stem from such bias is an essential direction for future research. This issue is especially important, given the relevance of intent to legal definitions of interpersonal discrimination (Washington v. Davis, 1976). More research is needed to understand how beliefs about the awareness of bias, controllability of its influence on behavior, and intention to engage in specific behaviors that turn out to be discriminatory shape perceptions of accountability for discrimination.

7.4. Conclusion

The present work revealed that professionals such as doctors and police officers are held less accountable for discriminatory behavior born of implicit, compared with explicit, bias. Moreover, this reduction in accountability was not accompanied by an increase in support for efforts to reduce bias through trainings or other institutional policies. Not only does this work have important implications for how researchers communicate the science of implicit bias, but it begs the question, can the discrimination born of implicit bias be reduced if no one is held responsible for it?

Open practices

The studies presented in this paper have earned Open Materials and Open Data badges for transparent practices. All data and materials for the studies presented here can be found at https://osf.io/5gtyv/.

Acknowledgements

This research was supported by an NSF Social, Behavioral, and Economic Sciences (SBE) Postdoctoral Research Fellowship (#1809370) awarded to the second author.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jesp.2019.04.010.

References

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin, 126*, 556–574. https://doi.org/10.1037//0033-2909.126.4.556.
Allport, G., W. (1954). The nature of prejudice. Oxford, England: Addison-Wesley.
Ames, D. L., & Fiske, S. T. (2013). Intentional harms are worse, even when they're not. *Psychological Science, 24*, 1755–1762. https://doi.org/10.1177/0956797613480507.
Axt, J. R., Ebersole, C. R., & Nosek, B. A. (2014). The rules of implicit evaluation by race, religion, and age. *Psychological Science, 25*, 1804–1815. https://doi.org/10.1177/0956797614543801.
Bargh, J. A. (1999). The cognitive monster: The case against the controllability of automatic stereotype effects. In S. Chaiken, & Y. Trope (Eds.). *Dual-process theories in social psychology* (pp. 361–383). New York, NY, US: Guilford Press.
Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review, 94*, 991–1013.
Cameron, C. D., Payne, B. K., & Knobe, J. (2010). Do theories of implicit race bias change moral judgments? *Social Justice Research, 23*, 272–289. https://doi.org/10.1007/s11211-010-0118-z.
Carter, E. R., & Murphy, M. C. (2015). Group-based differences in perceptions of racism: What counts, to whom, and why? *Social and Personality Psychology Compass, 9*, 269–280. https://doi.org/10.1111/spc3.12181.
Casad, B. J., Flores, A. J., & Didway, J. D. (2013). Using the implicit association test as an unconsciousness raising tool in psychology. *Teaching of Psychology, 40*, 118–123. https://doi.org/10.1177/0098628312475031.
Clack, Z. (Writer), Rhimes, S. (Credited Writer) & Sullivan, K. R. (Director). (January, 25, 2018). Personal Jesus [*Grey's Anatomy*]. In D. Allen (Executive Producer). B. Beers (Executive Producer). Z. Clack (Executive Producer). S. Collins (Supervising Producer). F. Einesman (Executive Producer). E. Finch (Co-Executive Producer). J. Ganzel (Co-Executive Producer). M. Gordon (Executive Producer). W. Harper (Executive Producer). M. Hope (Co-Producer). L. Klien (Supervising Producer). M. Marinis (Co-Executive Producer). E. Pompeo (Producer). A. Reaser (Co-Executive Producer). S. Rhimes (Creator). L. Taylor (Producer). K. Veroff (Executive Producer). S. E. White (Producer). Los Angeles, California: ABC Studios.
Cooley, E., Lei, R. F., & Ellerkamp, T. (2018). The mixed outcomes of taking ownership for implicit racial biases. *Personality and Social Psychology Bulletin, 44*, 1424–1434. https://doi.org/10.1177/0146167218769646.
Crandall, C. S., & Eshleman, A. (2003). A justification-suppression model of the expression and experience of prejudice. *Psychological Bulletin, 129*, 414–446. https://doi.org/10.1037/0033-2909.129.3.414.
Cushman, F. (2015). Deconstructing intent to reconstruct morality. *Current Opinion in Psychology, 6*, 97–103. https://doi.org/10.1016/j.copsyc.2015.06.003.
Dasgupta, N. (2004). Implicit ingroup favoritism, outgroup favoritism, and their behavioral manifestations. *Social Justice Research, 17*, 143–169. https://doi.org/10.1023/B:SORE.0000027407.70241.15.
Daumeyer, N. M., Rucker, J. M., & Richeson, J. A. (2017). Thinking structurally about implicit bias: Some peril, lots of promise. *Psychological Inquiry, 28*, 258–261. https://doi.org/10.1080/1047840X.2017.1373556.
Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56*, 5–18. https://doi.org/10.1037/0022-3514.56.1.5.
Devine, P. G., Monteith, M. J., Zuwerink, J. R., & Elliot, A. J. (1991). Prejudice with and without compunction. *Journal of Personality and Social Psychology, 60*, 817–830. https://doi.org/10.1037/0022-3514.60.6.817.
Devine, P. G., Plant, E. A., Amodio, D. M., Harmon-Jones, E., & Vance, S. L. (2002). The regulation of explicit and implicit race bias: The role of motivations to respond without prejudice. *Journal of Personality and Social Psychology, 82*, 835–848. https://doi.org/10.1037//0022-3514.82.5.835.
Dovidio, J. F., & Gaertner, S. L. (2010). *Intergroup bias.* (Handbook of Social Psychology).
Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology, 82*, 62–68. https://doi.org/10.1037//0022-3514.82.1.62.
Dunton, B. C., & Fazio, R. H. (1997). An individual difference measure of motivation to control prejudiced reactions. *Personality and Social Psychology Bulletin, 23*, 316–326. https://doi.org/10.1177/0146167297233009.
Eberhardt, J. L., Goff, P. A., Purdie, V. J., & Davies, P. G. (2004). Seeing black: Race, crime, and visual processing. *Journal of Personality and Social Psychology, 87*, 876–893. https://doi.org/10.1037/0022-3514.87.6.876.
Fiske, S. T. (2004). Intent and ordinary bias: Unintended thought and social motivation create casual prejudice. *Social Justice Research, 17*, 117–127. https://doi.org/10.1023/B:SORE.0000027405.94966.23.
Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132*, 692–731. https://doi.org/10.1037/0033-2909.132.5.692.
Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini meta-analysis of your own studies: Some arguments on why and a primer on how. *Social and Personality Psychology*

Compass, 10, 535–549. https://doi.org/10.1111/spc3.12267.
Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102*, 4–27.
Greenwald, A. G., & Pettigrew, T. F. (2014). With malice toward none and charity for some: Ingroup favoritism enables discrimination. *American Psychologist, 69*, 669–684. https://doi.org/10.1037/a0036056.
Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology. General, 143*, 1369–1392. https://doi.org/10.1037/a0035028.
Heider, F. (1958). The other person as perceiver. *The psychology of interpersonal relations* (pp. 59–78). Hillsdale, New Jersey: Lawrence Erlbaum Associates.
Hersh, E. D., & Goldenberg, M. N. (2016). Democratic and republican physicians provide different care on politicized health issues. *Proceedings of the National Academy of Sciences*, *113*, 11811–11816. https://doi.org/10.1073/pnas.1606609113.
Hester, N., & Gray, K. (2018). *For black men, being tall increases threat stereotyping and police stops. Proceedings of the National Academy of Sciences, 115, 2711–2715*. https://doi.org/10.1073/pnas.1714454115.
Kelly, D., & Roedder, E. (2008). Racial cognition and the ethics of implicit bias. *Philosophy Compass, 3*, 522–540. https://doi.org/10.1111/j.1747-9991.2008.00138.x.
Knobe, J., & Nichols, S. (2011). *Free will and the bounds of the self. In Oxford Handbook of Free Will (2nd ed., pp. 530–554). New York: Oxford University Press.* https://doi.org/10.1093/oxfordhb/9780195399691.003.0028.
Lagnado, D. A., & Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition, 108*, 754–770. https://doi.org/10.1016/j.cognition.2008.06.009.
Laurent, S. M., Nuñez, N. L., & Schweitzer, K. A. (2016). Unintended, but still blameworthy: The roles of awareness, desire, and anger in negligence, restitution, and punishment. *Cognition & Emotion, 30*, 1271–1288. https://doi.org/10.1080/02699931.2015.1058242.
Legault, L., Green-Demers, I., Grant, P., & Chung, J. (2007). On the self-regulation of implicit and explicit prejudice: A self-determination theory perspective. *Personality and Social Psychology Bulletin, 33*, 732–749. https://doi.org/10.1177/0146167206298564.
Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for behavioral sciences. *Behavior Research Methods, 49*, 433–442. https://doi.org/10.3758/s13428-016-0727-z.
Malle, B. F. (1999). How people explain behavior: A new theoretical framework. *Personality and Social Psychology Review, 3*, 23–48.
Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *The Journal of Psychology, 25*, 147–186. https://doi.org/10.1080/1047840X.2014.877340.
Mitchell, G., & Tetlock, P. E. (2006). Antidiscrimination law and the perils of mindreading. *Ohio State Law Journal, 67*, 1023–1122.
Monroe, A. E., & Malle, B. F. (2017). Two paths to blame: Intentionality directs moral information processing along two distinct tracks. *Journal of Experimental Psychology: General, 146*, 123–133. https://doi.org/10.1037/xge0000234.
Monroe, A. E., & Malle, B. F. (2019). People systematically update moral judgments of blame. *Journal of Personality and Social Psychology, 116*, 215–236. https://doi.org/10.1037/pspa0000137.
Monteith, M. J., Voils, C. I., & Ashburn-Nardo, L. (2001). Taking a look underground: Detecting, interpreting, and reacting to implicit racial biases. *Social Cognition, 19*, 395–417. https://doi.org/10.1521/soco.19.4.395.20759.
Nadler, J., & McDonnell, M.-H. (2012). Moral character, motive, and the psychology of blame. *Cornell Law Review, 97*, 255–304.
Nosek, B. A., Hawkins, C. B., & Frazier, R. S. (2011). Implicit social cognition: From measures to mechanisms. *Trends in Cognitive Science, 15*, 152–159.
Onyeador, I. N. (2017). *Presumed unintentional: The ironic effects of implicit bias framing on Whites' perceptions of discrimination (unpublished doctoral dissertation)*. Los Angeles, CA: University of California.
Payne, B. K., Vuletich, H. A., & Lundberg, K. B. (2017). The bias of crowds: How implicit bias bridges personal and systemic prejudice. *Psychological Inquiry, 28*, 233–248. https://doi.org/10.1080/1047840X.2017.1335568.
Payne, K., Niemi, L., & Doris, J. M. (2018). How to think about "implicit bias.". *Scientific American*. Retrieved from https://www.scientificamerican.com/article/how-to-think-about-implicit-bias/.
Perry, S. P., Murphy, M. C., & Dovidio, J. F. (2015). Modern prejudice: Subtle, but unconscious? The role of bias awareness in Whites' perceptions of personal and others' biases. *Journal of Experimental Social Psychology, 61*, 64–78. https://doi.org/10.1016/j.jesp.2015.06.007.
Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology, 75*, 811–832.
Plant, E. A., & Devine, P. G. (2009). The active control of prejudice: Unpacking the intentions guiding control efforts. *Journal of Personality and Social Psychology, 96*, 640–652. https://doi.org/10.1037/a0012960.
Redford, L., & Ratliff, K. A. (2016). Perceived moral responsibility for attitude-based discrimination. *British Journal of Social Psychology, 55*, 279–296. https://doi.org/10.1111/bjso.12123.
Shaver, K. G. (1985). *The attribution of blame: Causality, responsibility, and blameworthiness.* New York, NY: Springer-Verlag.
Simon, S., Moss, A. J., & O'Brien, L. T. (2019). Pick your perspective: Racial group membership and judgments of intent, harm, and discrimination. *Group Processes & Intergroup Relations, 22*, 215–232. https://doi.org/10.1177/1368430217735576.
Sommers, S. R., & Norton, M. I. (2006). Lay theories about white racists: What constitutes racism (and what doesn't). *Group Processes & Intergroup Relations, 9*, 117–138. https://doi.org/10.1177/1368430206059881.
Swim, J. K., Scott, E. D., Sechrist, G. B., Campbell, B., & Stangor, C. (2003). The role of intent and harm in judgments of prejudice and discrimination. *Journal of Personality and Social Psychology, 84*, 944–959. https://doi.org/10.1037/0022-3514.84.5.944.
Uhlmann, E. L., & Nosek, B. A. (2012). My culture made me do it: Lay theories of responsibility for automatic prejudice. *Social Psychology, 43*, 108–113. https://doi.org/10.1027/1864-9335/a000089.
Washington v. Davis, 426 U.S. (1976).
Weiner, B. (1995). *Judgments of responsibility: A foundation for a theory of social conduct.* New York: The Guilford Press.