Human Communication Research 0360-3989

ORIGINAL RESEARCH

How Communicating about Discrimination Influences Attributions of Blame and Condemnation

David C. DeAndrea and Olivia M. Bullock

School of Communication, The Ohio State University, Columbus, OH 43210, USA

Across two randomized experiments, we examine how communication about discriminatory acts can influence judgments of blame and condemnation. Specifically, we consider whether attributing discrimination to implicit or explicit bias affects how people evaluate online reports of discrimination. In Study 1 (N = 947), we explore this question in the context of an online news environment, and in Study 2 (N = 121) we replicate our results on a social media site (i.e., Twitter). Across both studies, we document how viewers respond differently to reports of discrimination due to variation in agent motives, the type of bias that purportedly caused the discriminatory behavior, and the extent to which agents are reported to have completed implicit bias training. We discuss our theoretical contribution to perspectives of blame attribution and the communication of bias as well as the practical implications of our findings.

Keywords: News Exposure, Mass Media, Social Media, Discrimination, Implicit Bias, Path Model of Blame, Attributions, Twitter, Morality, Socially Regulated Perspective of Blame

https://doi.org/10.1093/hcr/hqab016

Attribution refers to the cognitive process through which people seek to make sense of and explain social behavior and events (Jiang & Sun, 2020; Malle, 2011). Scholars have documented the importance of understanding how people make attributions (i.e., infer the cause of events) in many areas of communication, ranging from interpersonal, risk, and intercultural communication to media studies and persuasion (Bazarova & Hancock, 2010). Much of this work examines (a) how source and message features influence the attributions people make, and (b) how, once formed, attributions affect subsequent attitudes, beliefs, and actions. Consistent with this approach, a primary aim of this work is to explore how viewers interpret and respond to reports of discrimination online. Such work is needed because how instances of bias and discrimination are discussed in the news and on social media can

Corresponding author: David C. DeAndrea; e-mail: deandrea.1@osu.edu.

meaningfully influence public opinion and a willingness to engage in collective action (see Blanton & Ikizer, 2019; Dixon, 2020; Smith, Williamson, & Bigman, 2020).

A critical factor that can affect how people interpret and respond to online reports of discrimination is the type of bias that is reported to have caused the discriminatory act. Increasingly, media reports describe how acts of discrimination can occur due to people holding explicit or implicit biases (Blanton & Ikizer, 2019). This distinction is important for the evaluation of discriminatory behavior because the same discriminatory acts can receive less condemnation when they are described as being due to implicit rather than explicit biases (Daumeyer, Onyeador, Brown, & Richeson, 2019). Ultimately, this creates a unique challenge for communication scholars to explore. How can journalists, educators, and advocates discuss how different forms of bias cause discriminatory behavior without hampering efforts for reform and producing unintended consequences? Our first central aim is to directly address this question across two studies.

Beyond applying advances in attribution research to better understand the implications of communicating about the causes of discrimination online, we also test competing theoretical accounts of blame attribution. Whereas a socially regulated perspective of blame (e.g., the path model of blame; Malle, Guglielmo, & Monroe, 2014) suggests that people mainly adjust judgments of blame in an evenhanded manner based on relevant information (i.e., people are equally willing to absolve and condemn), motivated blame perspectives (e.g., Haidt, 2008) suggest that people engage in moral reasoning primarily to support their intuitions and are more willing to render and increase rather than lessen judgments of blame (Monroe & Malle, 2019). Therefore, a second central aim of this research is to examine the robustness of the socially regulated blame perspective in online environments, which scholars have speculated might encourage more biased blame attributions (Crockett, 2017; Monroe & Malle, 2019). Specifically, we test how people react to reports of discrimination in distinct online environments (Study 1: News Website; Study 2: Twitter) and inspect the moderating role of individual difference variables (i.e., political ideology and internal motivation to respond without prejudice).

Study 1

Communication and attributions of blame

Science communication and media effects scholars document how media reporting on controversial topics can meaningfully influence public opinion on issues of societal importance. For example, studies investigating unintended consequences of informing the public about science-related topics have found that communicating about climate change to skeptical groups can backfire (Hart & Nisbet, 2012). Similarly, it is possible that communication about implicit biases and discrimination might produce unintended results (Blanton & Ikizer, 2019). Indeed, recent

studies have examined how attributions of blame and condemnation for discriminatory acts can vary depending on whether reports indicate the acts were due to explicit or implicit biases (Daumeyer et al., 2019; Daumeyer, Onyeador, & Richeson, 2020). Although conceptual debates exist (see Greenwald & Lai, 2020; Payne, Vuletich, & Lundberg, 2017), implicit biases are generally defined as mental associations triggered automatically about a social category (e.g., race, gender) that can lead to discrimination without intent and possibly awareness (see Payne et al., 2017; Payne, Vuletich, & Brown-Iannuzzi, 2019). In contrast, people can self-report explicit biases upon introspection (Payne et al., 2019).

Attribution theories help explain why people would evaluate discriminatory acts differently depending on whether explicit or implicit biases spurred the prejudiced behavior. Specifically, the path model of blame (Malle et al., 2014) explains how, upon detecting a norm-violating act, people determine whether the transgressor intentionally committed the violation; judgments of intentionality increase blame whereas the the lack of intentionality mitigates blame. Although scholars have questioned whether implicit biases truly operate outside of conscious awareness, many researchers have used the terms implicit and unconscious bias interchangeably (see Payne et al., 2019). Furthermore, implicit biases are commonly described to the public as a form of unconscious bias that leads to discrimination without intent or awareness ; some articles even describe implicit biases as inherently unintentional.

Recently, Daumeyer et al. (2019, 2020) directly examined the implications of citing explicit or implicit bias as the cause of discrimination in news reports. In their first set of studies, the authors found that reports of various forms of discrimination (i.e., age, political affiliation, and racial/ethnic) produced less accountability for perpetrators, as well as less punishment, concern, and desire for reform, when they were described as caused by implicit rather than explicit biases (Daumeyer et al., 2019). These results subsequently replicated in follow-up studies which considered whether the observed relationships would generalize to the context of gender discrimination (Daumeyer et al., 2020). Together, this work provides robust evidence that discrimination attributed to implicit bias receives less condemnation, posing important implications for those who communicate about issues involving discrimination.

H1: People assign greater blame and condemnation for acts attributed to explicit biases relative to implicit biases.

We seek to extend upon this work in several ways. First, recent theoretical advances in blame attribution (i.e., the path model of blame) indicate the importance of going beyond a bifurcation of behavior into categories of intentional versus unintentional to fully explain how people make attributions of blame (see Malle et al., 2014; Monroe & Malle, 2019). From this perspective, within the category of intentional acts, people consider the reasons why the person performed the norm-violating action. Reasons can serve a blame-mitigating function (e.g., knocking someone over to prevent them from getting hit by a bus) or a blame-exacerbating

function (e.g., knocking someone over to cause them pain). For unintentional acts, people consider the degree to which an agent could have prevented the norm-violating action; preventability assessments involve whether the agent had the capacity and obligation to prevent the act. Greater perceived preventability exacerbates blame, whereas lower perceived preventability mitigates blame (Monroe & Malle, 2019).

The role of reasons

At first glance, the notion that there are "good" reasons for treating someone differently based on their race, gender, or sexual orientation might appear odd or repulsive. However, political spin often emphasizes different aspects of controversial events, news coverage and social media content can seek to justify the seemingly unjustifiable (e.g., the killing of Eric Garner; see Dukes & Gaither, 2017), and victim-blaming regularly occurs (Lumsden & Morgan, 2017; Paterson, Brown, & Walters, 2019). As such, it is necessary to explore how online reports can cite reasons to mitigate or exacerbate levels of blame and condemnation.

A faithful application of the path model of blame to the current context would suggest that online commenters can cite reasons to mitigate or exacerbate blame for transgressors who commit intentional acts of discrimination. However, the degree to which the presentation of reasons can exacerbate blame when online reports indicate that an agent intentionally discriminated against another individual or group is unknown. Intentional acts of discrimination lack ambiguity, which can standardize blame attributions (see Major et al., 2003). When no reasoning is presented for why an agent intentionally treated others differently based on, for instance, their race, it is easier for people to infer reasons (e.g., the agents want to harm the group) or causal history of reason explanations (e.g., the agents are racist) that exacerbate blame than it is to infer reasons that mitigate (i.e., lessen, not eliminate) blame (e.g., the agents believe they are helping the group). If negative reasons are highly accessible and inferred in absence of their explicit presentation, then we should not expect the presentation of such reasons to substantively increase levels of blame. Essentially, a ceiling effect might exist wherein it is difficult for the presentation of reasons to exacerbate blame for online reports that indicate an agent intentionally discriminated against another individual or group.

In general, considering the accessibility and valence of reasons for norm-violating acts can provide the same benefits outlined by those who advocate for an attribute-centered approach to understanding behavioral norms (see Rimal & Lapinski, 2015). Classifying behaviors by their attributes (e.g., public/private; frequent/infrequent) can illuminate when particular norms (e.g., descriptive, injunctive) will exert a substantive influence, increasing the predictive and explanatory power of social norm theories. Likewise, considering the accessibility and valence of reasons available to account for certain behaviors can be informative

for predicting how strongly reasons will exacerbate or mitigate blame. In the context of evaluating reports of intentionally discriminatory behavior online, we suspect the ability of reasons to mitigate blame is greater than the ability to exacerbate blame.

RQ1: Relative to providing no reasons, can the presentation of bad motives exacerbate blame and condemnation for intentional discriminatory acts?
H2: Relative to providing no reasons, the presentation of reasons that seek to justify intentional discriminatory acts mitigate blame and condemnation.

Preventability and implicit bias training

Similar to the role reasons play in mitigating or exacerbating blame for intentional acts, reporters, and online commenters can bring to the forefront the degree to which unintentional acts were preventable. When implicit biases are cited as the cause of discriminatory behavior, one factor that could affect perceptions of preventability is whether individuals have completed implicit bias training. Implicit bias training is regularly employed to reduce discriminatory behavior in a variety of professional and societal settings (Blanton & Ikizer, 2019; Sukhera, Watling, & Gonzalez, 2020). Although the efficacy of implicit bias training is complex (Bezrukova, Spell, Perry, & Jehn, 2016; Hagiwara, Dovidio, Stone, & Penner, 2020), and a recent meta-analysis indicates such training appears to produce trivial changes in explicit behavior (see Forscher et al., 2019), the public is regularly presented with implicit bias training as a viable option to reduce unintentional discrimination. The implications of discussing the presence or absence of implicit bias training for attributions of blame and the need for reform are understudied from a communicative perspective (Redfield, 2020).

Blanton and Ikizer (2019) discuss how science communication surrounding implicit bias tests and training might produce unintended consequences. Specifically, they note (a) how individuals might infer from media descriptions that they should inhibit implicit biases, but behavioral inhibition strategies often backfire, and (b) if implicit biases are increasingly described as being common, normative constraints that might otherwise hold such biases in check are reduced. We seek to examine the degree to which another mechanism (i.e., preventability perceptions) influences how people respond to reports of discrimination online. If implicit biases are reported as the cause of discrimination, specifying an unwillingness or failure to complete implicit bias training could exacerbate blame. Conversely, those seeking to mitigate blame might highlight efforts taken to reduce implicit biases through training completion.

H3: When discriminatory acts are attributed to implicit biases, information about the completion of implicit bias training mitigates blame (relative to no information).

H4: When discriminatory acts are attributed to implicit biases, information about the failure to complete implicit bias training exacerbates blame (relative to no information).

Potential moderating factors

Finally, we examine how individual difference variables might moderate blame attributions. As noted by Monroe and Malle (2019) and empirically documented (e.g., Schiller, Baumgartner, & Knoch, 2014), there are circumstances under which people are likely to attribute blame in a biased manner (e.g., favoring the ingroup and discriminating against an outgroup). The first individual difference that we investigate is internal motivation to respond without prejudice (IMS), which is defined as "the extent to which people are intrinsically motivated to behave in non-prejudiced ways" (Daumeyer et al., 2019). In their 2019 article, Daumeyer et al. (Study 3) investigate IMS as a potential moderator of the relationship between the type of information provided to participants and their blame judgments. Specifically, they considered whether those high in IMS may be especially likely to hold perpetrators of discriminatory acts accountable for their behavior, especially if the behaviors are grounded in explicit biases, whereas those low in IMS may be less likely to distinguish the blameworthiness of acts driven by explicit or implicit biases. Indeed, their results detected a moderation effect, suggesting that participants who were high in IMS generally tended to hold perpetrators more accountable for blameworthy acts. Those high in IMS were also more likely to support punishment in response to discriminatory acts due to explicit rather than implicit bias. Daumeyer and colleagues suggest the influence of IMS occurs because those low in IMS are less motivated to attend to details surrounding issues and information related to discrimination, whereas those high in IMS are more attentive to such information. The present study seeks to replicate and explore these relationships further.

RQ2: Does IMS influence judgments of blame and condemnation?

Lastly, we seek to extend previous work by examining another individual difference that may moderate how individuals make blame judgments: political ideology. There is clear evidence that public opinion on matters of discrimination varies greatly along partisan lines. For instance, in the United States, Democrats and Republicans have "vastly different views" on the treatment of Black Americans, the amount of attention given to racial issues, and their willingness to learn about racial inequities (Horowitz, Brown, & Cox, 2019, p. 8). There are also wide partisan gaps in views on discrimination in society based on gender, religion, and sexual orientation (Jones, 2019). Given the substantive differences in views on bias and discrimination in society, we seek to examine if political ideology influences judgments of blame and moral condemnation. Specifically, we consider whether political ideology influences people to make blame judgments in a manner less consistent with the socially regulated perspective of blame.

RQ3: Does political ideology influence judgments of blame and condemnation?

Method

Research design overview

A between-subjects experiment was designed to test the hypotheses and research questions. The bias information factor had six conditions that varied the type of bias reported (intentional only, intentional mitigating reason, intentional exacerbating reason, unintentional only, unintentional more preventable, unintentional less preventable). A second factor, agent type, had two conditions (individual or group) to test if the effects of the bias information factor generalized to both individuals and a group of individuals committing discriminatory acts. Thus, participants were randomly assigned to one of the 12 conditions.

Participants

A total of 1,009 participants were recruited through TurkPrime and paid $2.00 in exchange for their participation (see Litman, Robinson, & Abberbock, 2017). Although all participants received compensation, a modified version of a validated attention check by Berinsky, Margolis, and Sances (2014) was included that 62 participants failed; we subsequently removed these participants from all analyses yielding a final sample of 947. Sample characteristics are described in more detail in Supplementary Material.

Stimuli

Following the method employed by Daumeyer et al. (2019), participants read an online news article that varied information surrounding an act of discrimination. As a conceptual replication and extension, we changed the context surrounding the act to discriminatory lending, increased the number of bias information conditions, and varied whether the act of discrimination was undertaken by a single individual or a collection of individuals. In all conditions, participants read an online news article about the results of a bank audit investigating discriminatory lending practices. In all conditions, participants read that a report discovered that bank officer(s) approved loans for White applicants at almost double the rate of minority applicants, and that interviews with co-workers, and a review of video and audio recordings, confirmed the findings of the report. In the three intentional conditions (intentional only, mitigating reason, exacerbating reason), participants read that the investigation examined whether *explicit* biases caused discriminatory lending. The report always concludes that the agent(s) intentionally provided loans to White applicants at a greater rate than minority applicants. However, in the mitigating reason condition, the bank officer(s) explain that they are members of the local chamber of commerce, that they helped create a program to attract minority business owners in conjunction with the local credit union, and that they feel obliged to refer minority applicants to the program as it provides better interest rates. In the exacerbating reason condition, the bank officers(s) explain that they are members of the local chamber of commerce, and that it is within their rights to consider the types of

businesses and business owners they want to attract to the community and the types they feel do not fit within the fabric of the community.

In the three unintentional conditions (unintentional only, more preventable, less preventable), participants read that the investigation examined whether *implicit* biases caused discriminatory lending. The report always concludes that the difference in lending rates was unintentional and driven by implicit biases. However, in the more preventable condition, the report also states that the bank officer(s) failed to complete mandatory implicit bias training. In the less preventable condition, the report states that the bank officer(s) completed voluntary implicit bias training. Please see Supplementary Materials for all versions of the stimuli.

Measures

All items are measured on 7-point Likert style scales with endpoints of strongly disagree (1) to strongly agree (7) unless otherwise noted. In addition to measuring blame, multiple indicators of condemnation were included as outcome variables for two reasons. First, where possible we sought to replicate Daumeyer et al. (2019), who measured moral accountability and individual punishment. Second, Malle (2021) documents nuanced differences between classes of moral assessment that are reflected in our range of outcomes. In addition to measuring blame, we measured moral responsibility, agent favorability, and punishment perceptions as indicators of condemnation.

Blame

Following Monroe and Malle (2019), participants responded to the question, "How much blame does [agent] deserve?" using a slider bar with endpoints of 0–100.

Moral responsibility

Perceived moral responsibility was measured with eight items from Redford and Ratliff (2016). A sample item is "[The agent] is morally responsible for his actions" ($\alpha = .89$).

Agent favorability

Two sets of bipolar adjectives on 7-point scales were used to measure favorability perceptions. The stem read, "I view the [agent]. . ." and was followed by the bipolar adjectives unfavorably/favorably and negatively/positively (Spearman's rho = .96).

Individual punishment

Consistent with Daumeyer et al. (2019), we measured the degree to which individuals (rather than institutions) were perceived to deserve punishment. Participants read the stem, "The [agent] should be. . ." and responded to the following five items: reassigned, fined, suspended without pay, demoted, and fired ($\alpha = .90$).

Internal motivation to respond without prejudice subscale

The modified (Daumeyer et al., 2019) five-item subscale (Plant & Devine, 1998) was used to assess whether participants personally strive to act in nonbiased ways with others ($\alpha = .86$).

Conservativism

Two items were used to assess political ideology (Robinson, Shaver, & Wrightsman, 1999). The items stated "I endorse many aspects of conservative political ideology" and "I endorse many aspects of liberal political ideology" (reverse-coded; Spearman's rho = .53).

Results

Here, we organize the reporting of our results by dependent variable starting with the outcome of blame; we summarize support for each hypothesis in Table S11 of Supplementary Material. We first conducted a series of two-way ANOVAs to examine main and interaction effects of the bias information type and agent type factors for each outcome; A two-way ANOVA indicated a significant main effect for the bias information type factor, $F(5, 935) = 36.24$, $p < .001$, $\eta_p^2 = .16$, a non-significant main effect for the agent type factor, $F(1, 935) = .14$, $p = .71$, $\eta_p^2 < .001$, and a significant bias type x agent type interaction, $F(5, 935) = 2.54$, $p = .027$, $\eta_p^2 = .013$ on blame. Table 1 provides means and standard errors as well as an indication if the means significantly varied as determined by Bonferroni adjusted pairwise comparisons.

Hypothesis 1 states that people assign greater blame for acts attributed to explicit biases relative to implicit biases. Across both agent conditions, there were higher levels of blame in the intentional only ($M = 81.21$, $SD = 21.15$) condition relative to the unintentional only condition ($M = 57.82$, $SD = 29.43$; supporting H1). RQ1 explored whether, relative to providing no reasons, the presentation of bad reasons can exacerbate blame for intentional acts of discrimination. There was no difference between the intentional only and exacerbating reason conditions ($M = 80.33$, $SD = 21.14$), indicating that the presentation of bad reasons did not exacerbate blame. Hypothesis 2 predicted that, relative to providing no reasons, the presentation of reasons that seek to justify intentional acts of discrimination mitigate blame. The mitigating reason condition ($M = 52.84$, $SD = 33.42$) had significantly lower blame scores relative to both the intentional only and exacerbating reason conditions (supporting H2). Hypothesis 3 states that when discriminatory acts are attributed to implicit biases, information about the completion of implicit bias training mitigates blame, whereas Hypothesis 4 states information about the failure to complete implicit bias training exacerbates blame. The three unintentional conditions (i.e., unintentional only, less preventable, and more preventable) did not vary in blame within the collective agent condition (not supporting H3 and H4). However, the significant interaction effect provides some evidence that participants evaluated the

Table 1. Descriptive Statistics for Outcome by Information Condition

Outcome	Information Condition	Estimated Marginal Mean	Standard Error
Blame	Intentional Mitigating (a)	49.22[b, c, f]	3.06
Individual Agent	Intentional (b)	81.85[a, d, e, f]	2.98
	Intentional Exacerbating (c)	84.33[a, d, e, f]	2.98
	Unintentional Less Preventable (d)	52.37[b, c, f]	3.10
	Unintentional (e)	58.51[b, c]	2.96
	Unintentional More Preventable (f)	66.01[a, b, c, d]	2.96
Blame	Intentional Mitigating (a)	56.36[b, c]	3.02
Collective Agent	Intentional (b)	81.78[a, d, e, f]	2.96
	Intentional Exacerbating (c)	76.33[a, d, e, f]	2.98
	Unintentional Less Preventable (d)	59.19[b, c]	3.02
	Unintentional (e)	57.09[b, c]	3.04
	Unintentional More Preventable (f)	57.70[b, c]	2.96
Moral Responsibility	Intentional Mitigating (a)	4.36[b, c, f]	.10
	Intentional (b)	5.62[a, c, d, e, f]	.10
	Intentional Exacerbating (c)	5.57[a, d, e, f]	.10
	Unintentional Less Preventable (d)	4.49[b, c]	.10
	Unintentional (e)	4.63[b, c]	.10
	Unintentional More Preventable (f)	4.79[a, b, c]	.10
Favorability	Intentional Mitigating (a)	4.15[b, c, e, f]	.15
	Intentional (b)	2.54[a, c, d, e, f]	.14
	Intentional Exacerbating (c)	2.75[a, d, e, f]	.15
	Unintentional Less Preventable (d)	3.85[b, c]	.15
	Unintentional (e)	3.49[a, b, c]	.15
	Unintentional More Preventable (f)	3.44[a, b, c]	.14
Individual Punishment	Intentional Mitigating (a)	3.68[b, c]	.13
	Intentional (b)	5.09[a, c, d, e, f]	.12
	Intentional Exacerbating (c)	4.91[a, d, e, f]	.13
	Unintentional Less Preventable (d)	3.86 [b, c]	.13
	Unintentional (e)	3.85[b, c]	.13
	Unintentional More Preventable (f)	3.89[b, c]	.12

Note: Different letters in the same row indicate means that significantly differ at $p < .05$ after Bonferroni adjustment for multiple pairwise comparisons.

less preventable and more preventable conditions differently. Within the individual agent condition, the less preventable condition ($M = 52.36$, $SD = 26.76$) received significantly lower blame scores in comparison to the more preventable condition ($M = 66.01$, $SD = 25.06$; providing partial support for H3 and H4).

We repeated this analysis approach for the remaining outcome variables, which we collectively referred to as indicators of condemnation. However, given (a) the

interaction of the bias information and agent factors was not significant for the remaining dependent variables, and (b) the main effect of agent type was not significant for any of the remaining dependent variables, we focus our reporting on the main effects of the bias information factor.

A significant main effect of the bias information type factor was detected for the moral responsibility outcome, $F(5, 925) = 32.53$, $p < .001$, $\eta_p^2 = .15$. The intentional only condition and the exacerbating reason condition did not vary (informing RQ1) but had significantly higher scores for perceived moral responsibility relative to all other conditions (supporting H1 and H2). The mitigating reason condition did not differ from the less preventable condition but was significantly lower in perceived moral responsibility relative to the more preventable condition (lending minimal support to H3 and H4).

A significant main effect also emerged for the perceived favorability variable, $F(5, 935) = 18.12$, $p < .001$, $\eta_p^2 = .09$. Again, the intentional only and exacerbating reason conditions did not vary (RQ1) and were significantly lower in perceived favorability than all other conditions (supporting H1 and H2). The mitigating reason and less preventable condition did not vary; however, favorability perceptions were significantly higher in the mitigating condition relative to the unintentional only and more preventable conditions (lending minimal support to H3 and H4).

A main effect was also detected for individual punishment, $F(5, 933) = 24.38$, $p < .001$, $\eta_p^2 = .12$. Again, the intentional only and exacerbating reason conditions did not vary (RQ1) and were significantly higher in reported individual punishment than all other conditions (H1 and H2 supported); the remaining four conditions did not vary (no support for H3 or H4).

Next, we examined how conservatism and internal motivation to respond without prejudice (IMS) related to each dependent measure and whether conservatism and IMS moderated the effects of the bias information type factor on the dependent measures. As indicated by the results of multiple regression analyses presented in Table S3, conservatism and IMS independently predicted each outcome measure with one exception; conservatism did not significantly predict individual punishment scores when controlling for IMS. IMS positively predicted all outcome measures, except perceived favorability, with which it had a significant negative association. Higher scores on the conservatism measure were associated with lower scores for blame and moral responsibility and greater perceived favorability of the agent(s).

Model 1 in PROCESS (Hayes, 2018) was used to examine conservatism and IMS as moderators generating estimates of conditional effects based on 10,000 resamples. Given the multi-categorical nature of the bias information type factor, we examined if the difference between specific conditions (intentional only vs. unintentional only; exacerbating reason vs. mitigating reason; less preventable vs. more preventable) on each outcome varied as a function of conservatism or IMS. Tables S4–S6 provide the results of moderation analyses for IMS, and Tables S7–S9

show the results of moderation analyses for conservatism. We also visualize these results in Supplementary Materials.

IMS moderated the effect of the intentional only versus unintentional only condition on blame, moral responsibility, and favorability. The nature of the moderation varied (see Holbert and Park, 2020); contributory (blame, moral responsibility) and contingent (favorability) moderation was detected. For the contributory moderation, higher scores on IMS led to greater significant differences between the intentional only versus unintentional only conditions. For the contingent moderation, there was only a significant difference between conditions at moderate and high levels of IMS.

IMS also moderated the effect of the exacerbating reason versus mitigating reason condition on blame, moral responsibility, and individual punishment. All of the moderation effects were contributory in nature, with significant differences at all levels of IMS that increased in magnitude as IMS increased. There was not clear evidence of moderation by IMS for the less preventable versus more preventable condition for any of the outcomes.

Conservatism moderated the effect of the intentional only versus unintentional only condition on all outcomes. Contributory moderation was detected for blame, moral responsibility, and individual punishment, whereas contingent moderation was detected for perceived favorability. For the contributory moderation, the effect was significant at all levels of the moderator; however, the difference between conditions increased as conservatism scores increased. For the contingent moderation, there was not a significant difference between conditions at low levels of conservatism; however, there was a significant effect at moderate and high levels. Conservatism did not moderate the effect of the exacerbating reason versus mitigating reason conditions or the less preventable versus more preventable conditions on any outcome.

Discussion

The results indicate that the presentation of reasons and information about implicit bias training influence how people respond to reading about discrimination online. The exacerbating and intentional only conditions did not vary from one another but consistently led to more blame, moral responsibility, and punishment (and less agent favorability) than the other conditions. Overall support was found for H1 in that participants readily distinguished between intentional and unintentional acts. Whereas the mitigating reason condition reduced blame and condemnation (supporting H2), the exacerbating condition did not increase blame and condemnation relative to the intentional only condition (informing RQ1). For the outcome of blame, there was partial support for H3 and H4. The less preventable condition received significantly less blame than the more preventable condition, but only when individual targets were evaluated. However, for the other outcome measures, little support was obtained.

Addressing RQ2, IMS was independently associated with all dependent variables and moderated many of the relationships reported above. Greater IMS was associated with more blame and condemnation and greater differences between conditions. The moderation effects were primarily contributory in nature. Addressing RQ3, conservatism was significantly associated with most outcomes and consistently moderated the main effects between the intentional only and unintentional only conditions. The moderation effects were primarily contributory in nature.

Overall, the findings support a socially regulated perspective of blame. Participants consistently reduced blame and condemnation when intentional acts of discrimination were accompanied by a mitigating reason and when the same acts of discrimination were presented as unintentional and caused by implicit biases. Conservatism and IMS moderated some of the reported relationships; however, the moderation effects were primarily contributory in nature, indicating that participants all made significant adjustments in their levels of blame and condemnation, albeit of different magnitude. The data also support our speculation that the presentation of reasons might not be able to exacerbate blame above the amount people already assign for intentional acts of discrimination given the accessibility of bad motives for such acts. It is possible, however, that participants did not interpret the "exacerbating" reason presented in Study 1 as a bad motive. To address concerns about the unique effect that single messages for a specific act of discrimination might have had in Study 1, we sought to replicate our findings with multiple stimuli following the method of Monroe and Malle (2019).

Study 2

Study 2 expands upon Study 1 in the following ways. First, Study 1 examined reactions to one form of discrimination (racial/ethnic) in one context (banking). Study 2 follows the stimulus sampling approach of Monroe and Malle (2019) and examines the predictions and research questions outlined in Study 1 across reports of six different forms of discrimination (i.e., racial, gender/sex, sexual orientation, age, religious, and cultural) to increase the generalizability of our findings. Second, Monroe and Malle (2019) speculate that certain online environments (e.g., Twitter) might not possess the characteristics necessary for a socially regulated perspective to accurately account for blame attributions. They note, "Central to the socially regulated perspective is the claim that social demands for warrant motivate systematic moral information processing and judgments of blame. What follows from this contention is that intensifying or relaxing the requirement for warrant should modulate whether people are relatively more systematic or more biased in their judgments" (p. 233). We explore the degree to which the socially regulated perspective holds when evaluating accusations of discrimination in an online environment that potentially relaxes the need to justify or defend accusations of blame.

Method

Research design overview

Following the method of Monroe and Malle (2019), bias information type was a within-subjects factor with six conditions (intentional only, intentional mitigating reason, intentional exacerbating reason, unintentional only, unintentional more preventable, unintentional less preventable) and stimulus sampling was employed to enhance the generalizability of the results. We only included an individual agent version of each condition due to the results of Study 1.

Participants

A total of 126 participants were recruited through TurkPrime and paid $2.00 in exchange for their participation (see Litman et al., 2017). Participants from Study 1 were excluded from taking Study 2. Five participants failed the attention check for a final sample of 121 participants. Sample characteristics are available in Supplementary Material.

Procedure and stimuli

Participants were told that the researchers were interested in understanding how people react to events reported on Twitter and that they would view six different tweets (i.e., posts on Twitter) and provide their feedback on the events that were described. Participants were presented a series of six tweets that contained each bias information type. The source of each ostensible Tweet and reaction metrics (i.e., number of likes and retweets) were blacked out for all conditions. Six different events of discrimination (i.e., racial, gender/sex, sexual orientation, age, religious, and cultural) were created with unique versions for each bias information condition (intentional only, intentional mitigating reason, intentional exacerbating reason, unintentional only, unintentional more preventable, unintentional less preventable) for a total of 36 conditions (six of each topic × six of each information type). Participants were assigned to view six tweets containing each bias information condition in a randomly determined event and in a randomly determined order; participants never saw more than one version of each event or bias information type. See Supplementary Materials for the wording of all stimuli.

Measures

All items are measured on 7-point Likert scales with endpoints of strongly disagree (1) to strongly agree (7) unless otherwise noted. Whereas in Study 1 participants evaluated a single online news article, participants were required to evaluate six different tweets in Study 2 that described six different discriminatory events. As a result, it was necessary to adjust the number of items used to evaluate each event. Blame, agent favorability (Spearman's rho ranged from .88 to .97), IMS ($\alpha = .85$), and conservativism (Spearman's rho = .78) were measured in the same manner as in Study 1.

Moral responsibility

Perceived moral responsibility was measured with four items from Cameron, Payne, and Knobe (2010). The additional items added by Redford and Ratliff (2016) and used in Study 1 were dropped. Reliability estimates ranged from $\alpha = .72$ to .83.

Punishment

For the individual punishment measure, participants read the stem, "The [agent] should be. . ." and responded to the following four items: punished, suspended, demoted, and removed from their position. Reliability estimates ranged from $\alpha = .90$ to .97.

Results

We first conducted a series of repeated measures ANOVAs to examine the effect of the information condition on each outcome measure. When a main effect was detected, we employed Bonferroni adjusted pairwise comparisons to determine significant differences between the six information conditions (see Table 2). For the outcome of blame, a significant main effect was detected, Wilks' lambda = .38, $F(5, 115) = 37.90$, $p < .001$, $\eta_p^2 = .62$. Similar to Study 1, the intentional only and exacerbating conditions were significantly higher than all other conditions (supporting H1 and H2) and did not vary (RQ1). The mitigating reason, unintentional only, and less preventable conditions all had significantly lower blame scores than the more preventable condition, providing partial support for H3 and support for H4 (see Table 2).

A significant main effect was detected for the outcome of moral responsibility, Wilks' lambda = .47, $F(5, 114) = 25.60$, $p < .001$, $\eta_p^2 = .35$. Again, the intentional only and exacerbating conditions were significantly higher than all other conditions (supporting H1 and H2) and did not vary (RQ1). The mitigating and unintentional only conditions did not vary from the less preventable condition but were significantly lower in perceived moral responsibility than the more preventable condition (providing minimal support for H3 and support for H4).

A significant main effect was detected for perceived favorability, Wilks' lambda = .40, $F(5, 115) = 34.49$, $p < .001$, $\eta_p^2 = .60$. The intentional only and exacerbating conditions were significantly lower in perceived favorability relative to all other conditions (supporting H1 and H2). The more preventable condition was significantly lower than the mitigating, unintentional only, and less preventable conditions (providing partial support for H3 and support for H4).

A significant main effect was detected for the individual punishment measure, Wilks' lambda = .41, $F(5, 115) = 32.19$, $p < .001$, $\eta_p^2 = .59$. The intentional only and exacerbating conditions were significantly higher than all other conditions in individual punishment scores (supporting H1 and H2). The unintentional only condition did not vary from the less preventable condition but was significantly lower

Table 2. Study 2 Descriptive Statistics for Outcomes by Information Comparison

Outcome	Information Condition	Estimated Marginal Mean	Standard Error
Blame	Intentional Mitigating (a)	51.80[b, c, f]	3.15
	Intentional (b)	77.79[a, d, e, f]	2.57
	Intentional Exacerbating (c)	83.63[a, d, e, f]	2.17
	Unintentional Less Preventable (d)	55.08[b, c, f]	2.75
	Unintentional (e)	48.84[b, c, f]	2.59
	Unintentional More Preventable (f)	63.96[a, b, c, d, e]	2.73
Moral	Intentional Mitigating (a)	4.32[b, c, f]	.14
Responsibility	Intentional (b)	5.52[a, d, e, f]	.12
	Intentional Exacerbating (c)	5.78[a, d, e, f]	.11
	Unintentional Less Preventable (d)	4.60[b, c]	.12
	Unintentional (e)	4.36[b, c, f]	.11
	Unintentional More Preventable (f)	4.93[a, b, c, e]	.12
Favorability	Intentional Mitigating (a)	3.77[b, c, f]	.15
	Intentional (b)	2.47[a, d, e, f]	.14
	Intentional Exacerbating (c)	2.14[a, d, e, f]	.13
	Unintentional Less Preventable (d)	3.40[b, c, e, f]	.11
	Unintentional (e)	3.72[b, c, d, f]	.11
	Unintentional More Preventable (f)	3.00[a, b, c, d, e]	.13
Individual	Intentional Mitigating (a)	3.51[b, c]	.18
	Intentional (b)	4.89[a, d, e, f]	.16
Punishment	Intentional Exacerbating (c)	5.41[a, d, e, f]	.15
	Unintentional Less Preventable (d)	3.66[b, c]	.15
	Unintentional (e)	3.24[b, c, f]	.14
	Unintentional More Preventable (f)	3.94[b, c, e]	.16

Note: Different letters in the same row indicate means that significantly differ at $p < .05$ after Bonferroni adjustment for multiple pairwise comparisons.

in perceived moral responsibility than the more preventable condition (providing minimal support for H3 and support for H4).

Following Study 1, we next examined if IMS and conservatism moderated any of the above relationships. Model 2 in the MEMORE macro (Mediation and Moderation for Repeated Measures; Montoya & Hayes, 2017) was used to probe interaction effects. Given the multicategorical nature of the bias information type factor, we again examined if the difference between specific conditions (intentional only vs. unintentional only; exacerbating reason vs. mitigating reason; less preventable vs. more preventable) on each outcome varied as a function of IMS or conservatism. MEMORE calculates a difference score for each condition (e.g., intentional only blame score minus the unintentional only blame score) and then formally tests if this difference varies as a function of the moderator. In addition, MEMORE estimates the relationship between the moderator and each condition (e.g., regressing

intentional only blame score on IMS and regressing unintentional only blame score on IMS). Tables S12–S19 in Supplementary Materials provide the complete results of all moderation analyses.

IMS moderated the difference between the intentional versus unintentional conditions and the mitigating versus exacerbating conditions for all four outcome measures. A consistent pattern of contributory moderation was detected; higher IMS scores corresponded to a larger difference between conditions (see Tables S12–S15). For the intentional versus unintentional conditions, the significant moderation was more attributable to differences in the intentional conditions relative to the unintentional conditions. For the mitigating versus exacerbating conditions, the significant moderation was almost exclusively attributable to differences in the exacerbating conditions relative to the mitigating conditions. For the less preventable versus more preventable analyses, no moderation was detected. However, higher IMS scores in both the less preventable and more preventable conditions were associated with higher perceptions of moral responsibility and lower perceptions of agent favorability.

Conservatism only moderated the difference between the intentional only and unintentional only conditions for the outcome of moral responsibility; contributory moderation was detected such that as conservativism scores increased, the difference between scores for moral responsibility in the intentional only condition relative to the unintentional condition decreased. Although the *difference* between conditions (intentional vs. unintentional; mitigating vs. exacerbating; less preventable vs. more preventable) did not vary as a function of conservatism for almost every outcome, conservativism was significantly associated with several outcome measures. Specifically, conservativism was significantly related to every outcome for the intentional only, exacerbating, less preventable, and more preventable conditions. A consistent pattern emerged (see Tables S16–S19); higher scores of conservativism were associated with less blame and condemnation.

Discussion

Consistent with Study 1, the results of Study 2 indicate that the presentation of reasons and information about implicit bias training can influence how people respond to reading about discrimination on Twitter. Again, support was found for both H1 and H2; participants readily distinguished between intentional and unintentional acts, and the mitigating reason condition reduced blame, moral responsibility, and individual punishment while increasing the perceived favorability of the agent. As in Study 1, the exacerbating condition did not increase blame and condemnation relative to the intentional only condition (informing RQ1). There was minimal support for H3 and consistent support for H4 in this study. The results indicated that failing to complete mandatory implicit bias training (more preventable) increased blame and condemnation relative to the unintentional only condition. Although the less preventable condition and more preventable conditions significantly differed

for the outcomes of blame and favorability, the less preventable condition rarely differed from the unintentional only condition.

Addressing RQ2, the results were similar to Study 1. Higher scores of IMS corresponded with a larger difference between outcomes in the intentional and unintentional conditions and the exacerbating and mitigating conditions. Unlike Study 1, only one instance of moderation was detected for conservativism (RQ3). However, conservatism was significantly associated with many of the outcome measures in a similar manner as occurred in Study 1.

General discussion

Communicating about discrimination and bias

The collective results of our studies provide insight for how communicating about discrimination can affect perceptions of blame and condemnation. In online news articles and on social media, stating that discrimination occurred due to explicit or implicit biases meaningfully affected the amount of blame and condemnation that transgressors received. As outlined by the path model of blame, reasons were able to mitigate blame. However, as speculated, reasons were not able to exacerbate blame above and beyond the high levels already assigned when reports indicated that the discriminatory acts were intentional.

We chose to specifically examine how messaging about the completion (or lack thereof) of implicit bias training affects attributions of blame and condemnation when implicit biases are cited as the cause of discriminatory acts. Online sources regularly report on an array of organizations, companies, and government agencies that are seeking to reduce or eliminate discrimination due to implicit biases through training efforts. Overall, there was some evidence, particularly in Study 2, that blame and condemnation vary when information is presented about the completion of implicit bias training with stronger evidence that the failure to complete training increases blame rather than the completion of training lessening blame.

Both internal motivation to respond without prejudice (IMS) and conservatism were significantly associated in many instances with blame and condemnation. However, IMS demonstrated more robust moderation effects. Consistent with Daumeyer et al. (2019), higher IMS was associated with more blame and condemnation and a greater propensity to differentiate between discrimination due to implicit and explicit biases. Notably, the nature of the moderation was primarily contributory. The results of Study 1 indicated that conservatism moderated the effect of the intentional and unintentional conditions on outcomes, with evidence of contributory moderation for three outcomes and contingent moderation for one outcome. Study 2 indicated a lack of moderation by conservatism. Overall, the absence of moderation for significant effects and the detection of contributory moderation suggests that people across the political spectrum either *adjust* evaluations of blame and condemnation similarly based on intentionality information (no

moderation) or make similar, substantive adjustments that vary in magnitude (contributory moderation).

Theoretical and practical implications

It has been over a decade since communication scholars have explicitly advocated (Bazarova & Hancock, 2010) and empirically documented (DeAndrea & Walther, 2011) the benefits of using recent advances in attribution theory to illuminate communication phenomena. The folk conceptual theory of explanation (Malle, 1999) and the path model of blame (Malle et al., 2014) emphasize that people take into account decidedly different factors when evaluating intentional and unintentional behavior. Furthermore, the path model of blame provides nuanced explanations for how people evaluate behavior within these categories while avoiding problems that accompany alternative theoretical approaches (e.g., the covariation principle, person-situation distinctions; see Malle, 2011) often used by communication scholars (e.g., Tamborini et al., 2018). The current results indicate the value of these nuanced explanations; they elucidate how controversial topics of social importance can be communicated to the public in a manner that produces intended effects and minimizes unintended effects.

Path model of blame

Overall, the data provide robust support for a socially regulated perspective of blame that diverges from central elements of motivated blame perspectives (see Monroe & Malle, 2019). In general, there was consistent evidence that people were willing to minimize levels of blame and condemnation due to a lack of intentionality and the presentation of mitigating reasons. This evidence was obtained (a) for the evaluation of discriminatory acts which increase the potential for social desirability biases (see Perinelli & Gremigni, 2016), (b) in online settings anticipated to favor motivated blame perspectives, and (c) across individual differences that research suggests would minimize the likelihood of blame mitigation.

Although moderation effects were found for IMS across Study 1 and Study 2, and for conservatism in Study 1, a clear majority of these moderating effects were contributory in nature; participants significantly adjusted their blame and condemnation judgments due to the lack of intentionality or the presentation of mitigating reasons, but did so to different degrees. Digging deeper into the nature of the moderation effects provides additional illumination.

In Study 2, IMS moderated the magnitude of the differences between conditions. However, the moderation was primarily due to participants who varied in IMS providing different evaluations in the intentional only and exacerbating conditions. That is, participants who varied in IMS reduced blame and condemnation similarly in the unintentional only and mitigating conditions, whereas participants who varied in IMS differed in the level of blame they assigned to transgressors in the intentional only and exacerbating conditions. This suggests that all participants were willing to consider factors that mitigate blame; however, some participants (low in

IMS) reported lower levels of blame and condemnation than others (high IMS) for intentional acts of discriminatory behavior.

Furthermore, in Study 2, the way conservatism was associated with blame scores is consistent with a socially regulated perspective of blame. For instance, conservatism was associated with blame in the exacerbating condition with participants −1 *SD* below the mean having an average blame score of 92.53 and participants +1 *SD* above the mean having a score of 74.74. Conversely, in the mitigating condition, blame scores did not significantly vary by conservatism (−1 *SD* below the mean = 55.10; +1 *SD* above the mean = 48.49). Thus, blame mitigation consistently occurred across the political spectrum. The ceiling of blame judgments varied by conservatism but the gradedness of differences between conditions was similar.

That a socially regulated perspective of blame held under conditions anticipated to favor motivated blame perspectives helps extend the explanatory range and predictive scope of the path model of blame. However, our results also help establish a boundary condition that can increase the predictive accuracy of the path model of blame moving forward. The results help inform how the valence and accessibility of reasons commonly connected to specific actions can be used to explain and predict the degree to which reasons are likely to mitigate or exacerbate blame. Particularly in Study 2, a range of noxious reasons were cited as the cause of various forms of discrimination and yet participants reported an equivalently high level of blame and condemnation for intentional acts of discrimination presented without reasons.

Implications for science communication

Blanton and Ikizer (2019) encourage researchers to, ". . .apply exploratory and confirmatory methods to develop and test theories about both the intended and unintended influences that dominant science communications are having on public perception, with an eye toward ideological difference that might moderate effects" (p. 174). Following this directive, the current work documents how the path model of blame can be used by science communication scholars to predict and explain why science messages are more or less likely to produce intended and unintended effects. Through the lens of the path model of blame, identifying implicit biases as the cause of discriminatory behavior raises certain complexities—complexities that are not unique to the scientific communication of the causes of discrimination. If science communicators wish to educate the public on the causes that lead people or organizations to unintentionally produce negative outcomes, the path model of blame identifies ways to have nuanced discussions about such issues without reducing the culpability of transgressors.

For instance, *if* discrimination due to implicit bias is described as unintentional, communicating about the preventability of implicit biases can affect behavioral attributions and public opinion. In our work, a brief statement indicated whether or not agents completed some form of implicit bias training. Our subtle inductions led to some small effects on blame assessments wherein those who failed to complete required training received higher levels of blame and condemnation. Future work is

needed to further examine how implicit biases can be discussed as causes of discriminatory behavior in conjunction with emphasizing the capacity and obligation that agents have to overcome these biases, so as to not impede efforts of reform and produce unintended effects by reducing normative constraints (see Blanton & Ikizer, 2019). More generally, science communication and framing scholars can test messaging strategies that best convey to the public the capacity and obligation individuals have to prevent unintentional acts that contribute to negative societal outcomes.

Implications for mass media research

The scope and nature of framing is debated (see Cacciatore et al., 2016), yet there is a clear consensus that news media regularly employ episodic frames in their coverage (Boukes, 2021; Iyengar & Simon, 1993). Unlike thematic frames that discuss issues in a more general and abstract manner, episodic frames present specific, concrete events—similar to the reports across our two studies. Although researchers often compare episodic and thematic news frames, the current findings and the path model of blame can inform our understanding of how variation within episodic frames can meaningfully influence attributions of blame. Episodic frames can greatly influence attributions of responsibility (Boukes, 2021), with recent findings in the context of reporting on a financial crisis indicating that episodic frames can transfer responsibility away from the individual toward societal structures. Such findings run counter to expectations that individuals struggle to generalize from the specific (i.e., the episodic frame), thus keeping blame more at the individual level. The path model of blame provides the potential to organize conflicting findings in the literature on the effects of episodic frames; episodic frames that document intentionality, the absence of mitigating reasons, and/or the presence of exacerbating reasons should intensify the focus on the individual. In contrast, episodic frames that highlight the lack of intentionality and the inability or capacity to prevent negative outcomes might not only minimize attributions of blame, but also might encourage a focus on larger-scale causes of societal ills.

Our findings are also relevant to work that examines the effects of mass media depictions of minority groups (see Dixon, 2020). Recent work indicates how negative news coverage can inspire minority groups to engage in collective action due to increased collective efficacy (Saleem, Hawkins, Wojcieszak, & Roden, 2019). However, increased perceptions that the negative depictions of one's minority group were accurate muted these effects. The current results indicate how news coverage that emphasizes mitigating reasons, the lack of intentionality, or limited preventability can reduce blame and condemnation. If coverage includes justifications for events or policies that are discriminatory or portray minority groups negatively, viewers might be more inclined to view the hyperbolic, inaccurate, or one-sided coverage as an accurate snapshot of reality. Furthermore, if the negative coverage itself is presented, for instance, as fair and balanced, viewers might not recognize to the same degree the inaccurate and one-sided nature of the coverage.

Correspondingly, future work should seek to examine how news coverage that disproportionately focuses on mitigating factors or frames its coverage as fair can influence the degree to which viewers believe the coverage is negative and accurate, thus influencing their likelihood to engage in collective action.

Implications for new media research

Finally, our results have implications for studying how channel differences might influence the production and evaluation of moral condemnation. Scholars have suggested that features of online environments encourage the expression of moral outrage (see Crockett, 2017) and relax the need to justify accusations of blame and condemnation (Monroe & Malle, 2019). However, the results of Study 2 were largely consistent with Study 1 despite (a) the accusations of discrimination appearing on Twitter, and (b) the reduction of contextual information that justified the claims. Thus, the results reinforce that future work needs greater precision when determining how the production and evaluation of moral outrage meaningfully varies due to the channel through which it is expressed.

DeAndrea & Walther (2011) cautions communication technology scholars to avoid assuming qualities of digital media influence behaviors without engaging in careful channel comparisons that isolate features of media purported to produce an effect and examining their import. Huskey et al. (2018) make similar arguments directly related to examining the role that features of digital media might play in the production and evaluation of moral outrage online, suggesting more universal factors that might help account for what occurs on primarily text-based social media such as Twitter. The results of the current study bolster assertions that future work should seek to examine how, for instance, the salience of source cues, audience metrics, or perceived affordances (e.g., anonymity) of social media influence the production or evaluation of moral outrage, disentangling more universal human communication processes from communicative exchanges that are at least somewhat a byproduct of their medium. Deterministic claims that media (partially) drive specific behavior should be understood as such and directly tested. An array of individual and socio-cultural factors can interact with features of online environments to affect the processing of information online (see Weeks & Lane, 2020), and scholars have suggested intriguing places to start (e.g., Crockett, 2017; Huskey et al., 2018).

Limitations

A notable limitation of this work relates to our implicit bias training inductions. From a message effects perspective (see O'Keefe, 2003), our primary focus was to examine how including information about completing or failing to complete implicit bias training would influence attributions of blame; ensuring variability in preventability perceptions was not our primary goal. An alternative approach would have been to maximize the experimental variance in preventability perceptions. Future work should seek to examine how highlighting the completion (or not) of

implicit bias training can meaningfully alter perceptions of preventability directly or in conjunction with other message strategies. It is possible that all people are viewed as obligated to prevent certain acts, such as discrimination, and thus the capacity component of preventability assessments requires direct targeting. The relatively stronger support for the implicit bias training inductions in Study 2 is promising for future work given that the inductions in Study 1 were much more subtle.

Finally, future work should seek to disentangle the degree to which simply citing explicit or implicit biases as causes of discrimination produce differential attributions and evaluations without additional contextual information such as detailing awareness or intentionality. In popular press articles it is rare to see implicit biases cited without some reference that (a) people are unaware of such biases or (b) they produce unintentional forms of discrimination. Although contested, some scholars have even argued that people are not aware of their implicit attitudes and have used the terms implicit and unconscious bias interchangeably (see Payne et al., 2019). As our understanding of the nature of biases advances, researchers should continue to tease apart how media coverage evolves and produces corresponding effects on public opinion. We hope that the findings described here offer direction for future research and practical implications for the nuances of communicating about discriminatory acts.

Data availability statement

The data underlying this article will be shared on reasonable request to the corresponding author.

Supporting information

Additional Supporting Information may be found in the online version of this article.

Please note: Oxford University Press is not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

References

Bazarova, N. N., & Hancock, J. T. (2010). . In C. Salmon (Ed.), *Communication Yearbook* 34 (pp. 63–91). Mahwah, NJ: Erlbaum. https://doi.org/10.1080/23808985.2010.11679096

Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2014). Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *American Journal of Political Science*, 58(3), 739–753. https://doi.org/doi:10.1111/ajps.12081

Bezrukova, K., Spell, C. S., Perry, J. L., & Jehn, K. A. (2016). A meta-analytical integration of over 40 years of research on diversity training evaluation. *Psychological Bulletin*, 142(11), 1227–1274. https://doi.org/doi:10.1037/bul0000067

Blanton, H., & Ikizer, E. G. (2019). Elegant science narratives and unintended influences: An agenda for the science of science communication. *Social Issues and Policy Review*, 13(1), 154–181. https://doi.org/doi:10.1111/sipr.12055

Boukes, M. (2021). Episodic and thematic framing effects on the attribution of responsibility: The effects of personalized and contextualized news on perceptions of individual and political responsibility for causing the economic crisis. *The International Journal of Press/Politics. Advance online publication.* https://doi.org/10.1177/1940161220985241

Cacciatore, M. A., Scheufele, D. A., & Iyengar, S. (2016). The end of framing as we know it. . . and the future of media effects. *Mass Communication and Society*, 19*(*1*)*, 7-23. https://doi.org/10.1080/15205436.2015.1068811

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

Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour*, 1, 769–771. https://doi.org/doi:10.1038/s41562-017-0213-3

Daumeyer, N. M., Onyeador, I. N., Brown, X., & Richeson, J. A. (2019). Consequences of attributing discrimination to implicit vs. explicit bias. *Journal of Experimental Social Psychology*, 84, 103812. https://doi.org/doi:10.1016/j.jesp.2019.04.010

Daumeyer, N. M., Onyeador, I. N., & Richeson, J. A. (2020). Does shared gender group membership mitigate the effect of implicit bias attributions on accountability for gender-based discrimination? *Personality and Social Psychology Bulletin, 014616722096530.* https://doi.org/10.1177/0146167220965306

DeAndrea, D. C., & Walther, J. B. (2011). Attributions for inconsistencies between online and offline self-presentations. *Communication Research*, 38, 805–825. https://doi.org/doi:10.1177/0093650210385340

Dixon, T. (2020). Media stereotypes: Content, effects, and theory. In M. B. Oliver, A. A. Raney, & J. Bryant (Eds.), *Media effects: Advances in theory and research* (4th ed., pp. 243–257). Routledge.

Dukes, K. N., & Gaither, S. E. (2017). Black racial stereotypes and victim blaming: Implications for media coverage and criminal proceedings in cases of police violence against racial and ethnic minorities. *Journal of Social Issues*, 73(4), 789–807. https://doi.org/10.1111/josi.12248

Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2019). A meta-analysis of procedures to change implicit measures. *Journal of personality and social psychology*, 117*(*3*)*, 522-559. http://dx.doi.org/10.1037/pspa0000160

Greenwald, A. G., & Lai, C. K. (2020). Implicit social cognition. *Annual Review of Psychology*, 71, 419–445. https://doi.org/10.1146/annurev-psych-010419-050837

Hagiwara, N., Dovidio, J. F., Stone, J., & Penner, L. A. (2020). Applied racial/ethnic healthcare disparities research using implicit measures. *Social Cognition*, 38(Supplement), s68–s97. https://doi.org/10.1521/soco.2020.38.supp.s68

Haidt, J. (2008). Morality. *Perspectives on Psychological Science*, 3(1), 65–72. https://doi.org/10.1111/j.1745-6916.2008.00063.x

Hart, P. S., & Nisbet, E. C. (2012). Boomerang effects in science communication: How motivated reasoning and identity cues amplify opinion polarization about climate

mitigation policies. *Communication Research*, 39(6), 701–723. https://doi.org/10.1177/0093650211416646

Hayes, A. F. (2018). Introduction to mediation, moderation, and conditional process analysis: A regression based approach (2nd ed.). NewYork, NY: Guilford Press.

Holbert, R. L., & Park, E. (2020). Conceptualizing, organizing, and positing moderation in communication research. *Communication Theory*, 30(3), 227–246. https://doi.org/10.1093/ct/qtz006

Horowitz, J. M., Brown, A., & Cox, K. (2019). *Race in America 2019*. Pew Research Center. Retrieved from https://www.pewsocialtrends.org/2019/04/09/race-in-america-2019/

Huskey, R., Bowman, N., Eden, A., Grizzard, M., Hahn, L., Lewis, R., . . . Weber, R. (2018). Things we know about media and morality. *Nature Human Behavior*, 2, 315. https://doi.org/10.1038/s41562-018-0349-9

Iyengar, S., & Simon, A. (1993). News coverage of the Gulf crisis and public opinion: A study of agenda-setting, priming, and framing. *Communication research*, 20(3), 365–383. https://doi.org/10.1177/009365093020003002

Jiang, L. C., & Sun, M. (2020). Attribution. *The International Encyclopedia of Media Psychology*, 1–6. https://doi.org/10.1002/9781119011071.iemp0292

Jones, B. (2019). *Democrats far more likely than Republicans to see discrimination against blacks, not whites*. Pew Research Center. Retrieved from https://www.pewresearch.org/fact-tank/2019/11/01/democrats-far-more-likely-than-republicans-to-see-discrimination-against-blacks-not-whites/

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, 49(2), 433–442. https://doi.org/10.3758/s13428-016-0727-z

Lumsden, K., & Morgan, H. (2017). Media framing of trolling and online abuse: Silencing strategies, symbolic violence, and victim blaming. *Feminist Media Studies*, 17(6), 926–940. https://doi.org/10.1080/14680777.2017.1316755

Malle, B. (2011). Time to give up the dogmas of attribution: An alternative theory of behavior explanation. In J. M. Olson & M. P. Zanna (Eds.), *Advances in experimental social psychology* (Vol. 44, pp. 297–352). Burlington, VT: Academic Press

Major, B., Quinton, W. J., & Schmader, T. (2003). Attributions to discrimination and self-esteem: Impact of group identification and situational ambiguity. *Journal of Experimental Social Psychology*, 39(3), 220-231. https://doi.org/10.1016/S0022-1031(02)00547-4

Malle, B. F. (1999). How people explain behavior: A new theoretical framework. *Personality and social psychology review*, 3(1), 23-48. https://doi.org/10.1207/s15327957pspr0301_2

Malle, B. F. (2021). Moral judgments. *Annual Review of Psychology*, 72(1), annurev-psych-072220-104358. https://doi.org/10.1146/annurev-psych-072220-104358

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

Monroe, A. E., & Malle, B. F. (2019). People systematically update moral judgments of blame. *Journal of Personality and Social Psychology*, 116(2), 215–236. https://doi.org/10.1037/pspa0000137

Montoya, A. K., & Hayes, A. F. (2017). Two-condition within-participant statistical mediation analysis: A path-analytic framework. *Psychological Methods*, 22(1), 6–27. https://doi.org/10.1037/met0000086

O'Keefe, D. J. (2003). Message properties, mediating states, and manipulation checks: Claims, evidence, and data analysis in experimental persuasive message effects research. *Communication Theory*, 13*(3)*, 251-274. https://doi.org/10.1111/j.1468-2885.2003.tb00292.x

Paterson, J. L., Brown, R., & Walters, M. A. (2019). The short and longer term impacts of hate crimes experienced directly, indirectly, and through the media. *Personality and Social Psychology Bulletin*, 45(7), 994–1010. https://doi.org/10.1177/0146167218802835

Payne, B. K., Vuletich, H. A., & Brown-Iannuzzi, J. L. (2019). Historical roots of implicit bias in slavery. *Proceedings of the National Academy of Sciences*, 116(24), 11693–11698. https://doi.org/10.1073/pnas.1818816116

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(4), 233–248. https://doi.org/10.1080/1047840X.2017.1335568

Perinelli, E., & Gremigni, P. (2016). Use of social desirability scales in clinical psychology: A systematic review. *Journal of Clinical Psychology*, 72(6), 534–551. https://doi.org/10.1002/jclp.22284

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of personality and social psychology*, 75*(3)*, 811-832. https://doi.org/10.1037/0022-3514.75.3.811

Redfield, S. (2020). *Implicit bias is real, implicit bias training matters: Responding to the negative press*. SSRN. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3686762

Redford, L., & Ratliff, K. A. (2016). Perceived moral responsibility for attitude-based discrimination. *British Journal of Social Psychology*, 55(2), 279–296. https://doi.org/10.1111/bjso.12123

Rimal, R. N., & Lapinski, M. K. (2015). A re-explication of social norms, ten years later. *Communication Theory*, 25(4), 393–409. https://doi.org/10.1111/comt.12080

Robinson, J. P., Shaver, P. R., & Wrightsman, L. S. (Eds.). (1999). *Measures of political attitudes* (Vol. 1–2). Academic Press.

Saleem, M., Hawkins, I., Wojcieszak, M. E., & Roden, J. (2019). When and how negative news coverage empowers collective action in minorities. *Communication Research*. Online First. https://doi.org/10.1177/0093650219877094

Schiller, B., Baumgartner, T., & Knoch, D. (2014). Intergroup bias in third-party punishment stems from both ingroup favoritism and outgroup discrimination. *Evolution and Human Behavior*, 35(3), 169–175. https://doi.org/10.1016/j.evolhumbehav.2013.12.006

Smith, M. A., Williamson, L. D., & Bigman, C. A. (2020). Can social media news encourage activism? The impact of discrimination news frames on college students' activism intentions. *Social Media + Society*, 6(2), 205630512092136. https://doi.org/10.1177/2056305120921366

Sukhera, J., Watling, C. J., & Gonzalez, C. M. (2020). Implicit bias in health professions: From recognition to transformation. *Academic Medicine*, 95(5), 717–723. https://doi.org/10.1097/ACM.0000000000003173

Tamborini, R., Grall, C., Prabhu, S., Hofer, M., Novotny, E., Hahn, L., . . . Sethi, N. (2018). Using attribution theory to explain the affective dispositions of tireless moral monitors toward narrative characters. *Journal of Communication*, 68, 842–871. https://doi.org/10.1093/joc/jqy049

Weeks, B. E., & Lane, D. S. (2020). The ecology of incidental exposure to news in digital media environments. *Journalism*, 21, 1119–1135. https://doi.org/10.1177%2F1464884920915354