PNAS RESEARCH ARTICLE | SOCIAL SCIENCES OPEN ACCESS

# People who share encounters with racism are silenced online by humans and machines, but a guideline-reframing intervention holds promise

Cinoo Lee[a,b,1,2], Kristina Gligorić[b,c,1,2], Pratyusha Ria Kalluri[c,1], Maggie Harrington[a,1], Esin Durmus[c], Kiara L. Sanchez[d], Nay San[b,e], Danny Tse[c], Xuan Zhao[b], MarYam G. Hamedani[b], Hazel Rose Markus[a,b], Dan Jurafsky[c,e,2], and Jennifer L. Eberhardt[a,b,f,2]

Affiliations are included on p. 11.

Contributed by Jennifer L. Eberhardt; received December 27, 2023; accepted April 27, 2024; reviewed by Christopher A. Bail and Maarten Sap

**Are members of marginalized communities silenced on social media when they share personal experiences of racism? Here, we investigate the role of algorithms, humans, and platform guidelines in suppressing disclosures of racial discrimination. In a field study of actual posts from a neighborhood-based social media platform, we find that when users talk about their experiences as targets of racism, their posts are disproportionately flagged for removal as toxic by five widely used moderation algorithms from major online platforms, including the most recent large language models. We show that human users disproportionately flag these disclosures for removal as well. Next, in a follow-up experiment, we demonstrate that merely witnessing such suppression negatively influences how Black Americans view the community and their place in it. Finally, to address these challenges to equity and inclusion in online spaces, we introduce a mitigation strategy: a guideline-reframing intervention that is effective at reducing silencing behavior across the political spectrum.**

content moderation | social media | natural language processing (NLP) | race | toxicity classification

## Significance

Content moderation practices on social media risk silencing voices of historically marginalized groups. We find that posts in which users share personal experiences of racism are disproportionately flagged by both algorithms and humans. Not only does this hinder the potential of social media to give voice to marginalized communities, we also find that witnessing such suppression could exacerbate feelings of isolation, both online and offline. We offer a path to reduce flagging among users through a psychologically informed reframing of moderation guidelines. In an increasingly diverse nation where online interactions are commonplace, these findings highlight the need to foster more productive and inclusive conversations about race-based experiences and we demonstrate how content moderation practices can help or hinder this effort.

The widespread adoption of the internet has fundamentally altered how people engage with one another, with social media platforms emerging as pivotal arenas for social interaction and discourse (1). Unlike traditional face-to-face interactions, social media platforms offer a medium where even minority voices sharing their personal experiences and perspectives can reach a broad audience. Recent social justice movements, such as #livingwhileblack and #metoo, show how social media can provide historically marginalized communities with a public platform to discuss their experiences with discrimination and inequality in ways that create social impact (2).

Concurrently, both social media companies and lawmakers have been increasingly focused on addressing issues such as online harassment, toxicity, and hate speech. Marginalized groups often disproportionately bear the brunt of these negative online experiences (3, 4). Social media platforms have instituted guidelines outlining acceptable content and behavior and enforce these guidelines through content moderation practices, which rely on natural language processing (NLP) algorithms, human oversight, or a combination of both (5–8). Automated algorithms excel at efficiently managing large volumes of data, whereas human reviewers can provide a more nuanced grasp of the social context (9).

Although content moderation practices aim to create safe and inclusive online environments, there is growing concern that these efforts may, paradoxically, discriminate against marginalized voices (10, 11). Content created by users from marginalized groups, for example, can face unwarranted removal even when they do not violate platform guidelines or create harm. One plausible cause for such removal is that when people share their perspectives and racialized experiences online, content moderation algorithms may struggle to discern the difference between race-related talk and racist talk (12). Moreover, human reviewers may opt to remove race-related content, deeming such content uncomfortable, inappropriate, or contentious (13–16).

Individuals whose content is marked for removal face more than just content loss. Multiple flags could lead to account suspension, isolating people from their social networks and resources, sometimes jeopardizing the missions and livelihoods of small business owners and nonprofits reliant on these platforms for daily operations (17, 18). In fact, Instagram users whose online activity suggested they were Black were about 50% more likely to be subjected to automatic account suspension by the moderation system, compared to their White counterparts (19). Similarly, Black Facebook users have reported being silenced when discussing racism on the platform, resulting in account suspension

Competing interest statement: J.L.E. serves on the advisory board for a social media company. The position is unpaid.



[1]C.L., K.G., P.R.K., and M.H. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: cinoolee@stanford.edu, gligoric@stanford.edu, jurafsky@stanford.edu, or jleberhardt@stanford.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2322764121/-/DCSupplemental.

Published September 9, 2024.

for weeks or even months (17). Criticism has also been directed at TikTok algorithms for disadvantaging and banning Black creators' work (20, 21). Developing content moderation practices that ensure online community safety without perpetuating bias is a vital step toward realizing the positive potential of social media.

Here, we examine the roles that automated content moderation tools, human moderation behaviors, and platform moderation guidelines play in suppressing the voices of people of color and the potential consequences of these widespread moderation practices. We define suppression in this context as a post being labeled toxic by algorithms, or being flagged by humans; both actions can result in content being down-ranked or removed from a platform, preventing other users from seeing or engaging with it. Despite researchers identifying flaws in toxicity algorithms and social media users from marginalized groups consistently reporting instances of suppression in both survey research (22) and mass media (17–20), there is a lack of empirical research testing these observations by utilizing actual flagging and toxicity rates within real embedded systems.

We center our attention on racial discrimination disclosures: instances where individuals from historically marginalized racial groups share their own or their close others' experiences with discrimination and inequality (see Fig. 1*A* and *SI Appendix* for examples). Self-disclosures such as these can not only foster empathy, social support, and stronger connections (23, 24), but can also serve as a means to address grievances and potentially drive tangible change (25). Previous research has shown that harm-related personal experiences can even help bridge moral and political divides (26). Here, we theorize that racial discrimination disclosures could increase community awareness and understanding by shedding light on the frequency and severity of discrimination faced by people of color.

However, existing research also documents negative consequences of disclosing racial discrimination in interpersonal interactions. For example, disclosers may be viewed as agitators or troublemakers (27–29). Online, this particular form of race talk may be unnecessarily flagged for removal: Algorithms may flag racial discrimination disclosures due to the negative nature of the experiences described, while users may penalize authors of this content for challenging dominant narratives of colorblindness and racial progress (30, 31).

While self-disclosure on social media has become common in recent years, a comprehensive analysis of racial discrimination disclosures in particular has not yet been conducted. As U.S. American neighborhoods grow increasingly diverse (32, 33), and as opportunities to engage increase with the rise of social media, understanding how communities grapple with conversations about timely and complex topics such as race becomes more vital. Thus, we investigate flagging of racial discrimination disclosures as one impactful form of harm people of color face on social media.

Our research investigates whether current publicly available out-of-the-box algorithms used for online content moderation silence the voices of historically marginalized racial groups, and if so, what linguistic features of content influence flagging behavior (Studies 1a and 1b). We then compare how humans respond to racial discrimination disclosures, and what psychological processes influence human flagging decisions (Studies 2a and 2b). We establish the impact of such suppression on marginalized members of the community through an online experiment (Study 3). Finally, we test an intervention for reducing flagging behavior by reframing conventional platform moderation guidelines (Study 4).

## Data: Racial Discrimination Disclosures

To our knowledge, there is no existing dataset of racial discrimination disclosures. Given this, we compiled and leveraged a large dataset of posts from a social media platform that aims to build connections among people in their local communities. We aimed to better understand how users from marginalized groups discuss the racial discrimination they experience in their day-to-day lives and how content moderation might play a role in stifling those discussions.

Using a combination of computational and manual annotation, we identified 1,025 racial discrimination disclosures shared across 44 states in the United States. (*Materials and Methods*). We defined racial discrimination disclosures as sharing of negative experiences related to race that were experienced by the poster or

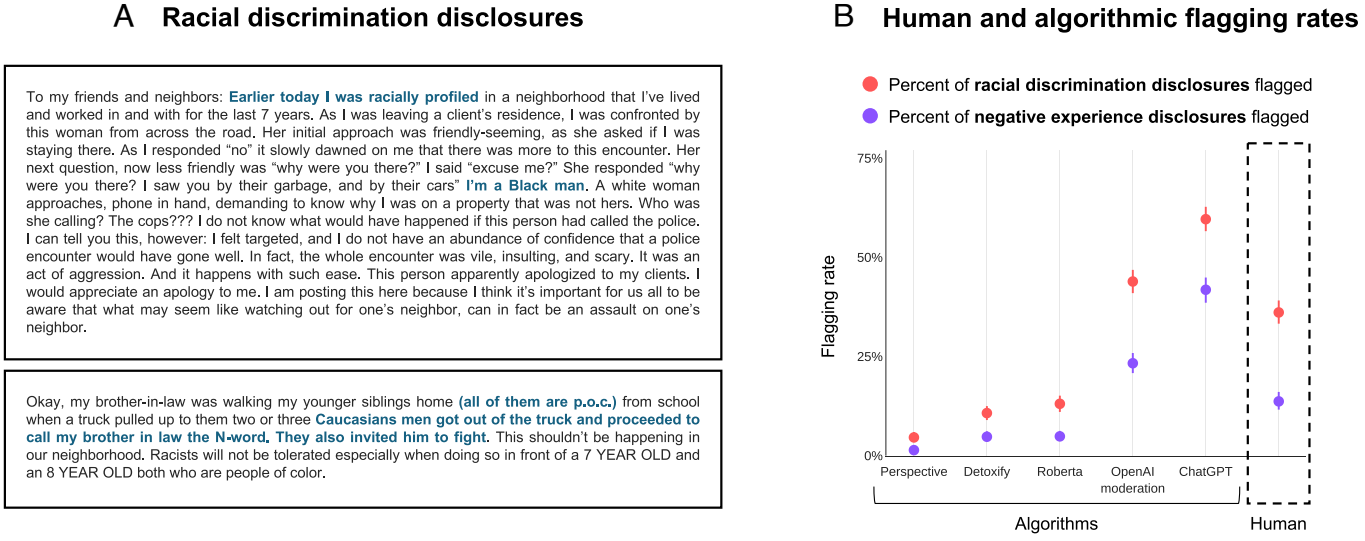


**Fig. 1.** (*A*) Examples of racial discrimination disclosures shared online (the posts have been slightly redacted for anonymity and brevity). (*B*) Rates of algorithmic and human flagging (for removal), displayed for racial discrimination disclosures and negative interpersonal experience disclosures, respectively. Error bars indicate SE of the mean.

close others (such as a spouse, friend, or child). These experiences could be a specific event or could refer to a pattern of racism that the poster or close other had experienced over time. We also established a control group of 1,009 negative interpersonal experience disclosures, which are posts that describe negative experiences of the poster or a close other that do not mention or indicate race (see *Materials and Methods* for annotation details and *SI Appendix*, section 1.C for coding criteria and examples); this dataset allows us to control for both emotional valence and type of content (i.e., sharing of negative interpersonal experiences).

## Study 1a: Do Algorithms Flag Racial Discrimination Disclosures as Toxic?

Content moderation algorithms have become ubiquitous across social media platforms, offering speed, scalability, and the consistent application of predefined rules. However, their influence on dialogs concerning racial discrimination and inequality remains an unanswered question—do these algorithms encourage or obstruct such discourse? We conducted an assessment of prominent publicly available off-the-shelf classifiers used for toxicity detection. These systems are developed and utilized by leading tech companies: OpenAI moderation Application Programming Interface (API) (OpenAI), Perspective API (Google), Roberta (Facebook), and Detoxify (Unitary, an online content moderation company). Given recent interest in large language models for content moderation tasks (34, 35), we also include results from ChatGPT (OpenAI's GPT-4), generated through prompting for toxicity detection. Our focus was on how algorithms assess the toxicity of racial discrimination disclosures. In the context of online discussions, Perspective API, Roberta, and Detoxify define toxicity as "rude, disrespectful, or unreasonable [content] that is likely to make someone leave a discussion."*

In our evaluation (*Materials and Methods*), we compared how these five modern toxicity detectors rated the toxicity of racial discrimination disclosures ($N = 1{,}025$), compared to negative interpersonal experience disclosures ($N = 1{,}009$). Across all tested algorithms, racial discrimination disclosures were flagged as toxic significantly more than negative interpersonal experience disclosures, with flagging rates ranging from 4.59% (Perspective API) to 59.61% (ChatGPT), compared to rates of 1.39% (Perspective API) to 41.82% (ChatGPT) for the negative interpersonal experience disclosures (Fig. 1*B*). While there is variation in flagging rates across algorithms, all five models are more likely to flag racial discrimination disclosures as toxic than negative interpersonal experience disclosures, despite the two datasets being comparable on attributes such as negative emotion [as defined by Vader (36)] and profanity (defined as the swearwords category in Linguistic Inquiry and Word Count (LIWC) lexicons (37); see details in *SI Appendix*, section 1.F). Prior research has documented that earlier content moderation algorithms misclassify mere mentions of minority identity as toxic (10, 38), and that social media platforms tend to label discussion of racism as hate speech (22, 39–41). While efforts exist to debias toxicity detection models (42), we found that disproportionate flagging of racial discrimination disclosures persist even when testing with debiased models and are challenging to address through data augmentation (*SI Appendix*, section 2.C). Despite significant recent improvements in language processing technology, our work shows that even the latest content moderation systems misclassify personal narratives by victims of racism as toxic. These systematic flaws can lead to disproportionate removal of online content produced by historically marginalized groups.

*https://perspectiveapi.com/how-it-works/.

## Study 1b: What Predicts Algorithmic Flagging of Racial Discrimination Disclosures as Toxic?

What may drive algorithmic flagging of racial discrimination disclosures? Recent work suggests that content moderation algorithms are oversensitive to certain language cues, such as dialectic markers or identity mentions (10, 38, 43–45). We extend this work to hypothesize that algorithmic flagging decisions might have trouble comprehending nuanced language because they overrely on lexical markers of affect, potentially overlooking the broader context in which these affective markers are embedded.

To examine this, we extracted lexical markers of affect (positive emotion, negative emotion, and profanity) and fitted logistic regression models to measure the impact of these affective markers on algorithmic flagging of racial discrimination disclosures (*Materials and Methods*). The full regression modeling statistics are listed in *SI Appendix*, section 3.E.

**Limitation of Algorithmic Flagging: Affective Lexical Cues.** Consistent with our hypothesis, affective markers significantly influence algorithmic flagging across all models (Fig. 2*A*). The presence of positive emotion words tended to decrease the likelihood of algorithmic flagging, while the presence of negative emotion words and swear words increased flagging probabilities. These results stand in contrast with human flagging behavior, where the inclusion of positive emotion words and profanity did not significantly impact flagging rates of racial discrimination disclosures (*SI Appendix*, section 3.C). This suggests that algorithms may be particularly susceptible to affective lexical cues, while potentially overlooking contextual nuances that should influence their interpretation (46). Algorithms struggled to differentiate whether a swear word was used as part of a user's language or merely quoted within a description of a discriminatory remark faced by the user—a nuance that is discernible by human readers (*SI Appendix*, section 3.A).

We note that our racial discrimination disclosure and negative experience disclosure datasets contain similar levels and frequencies of negative emotion and swear words (*SI Appendix*, section 1.F) and that these affective markers also influence algorithmic flagging of negative experience disclosure, though the effect sizes are much smaller. This suggests that while algorithms may be overly influenced by affective lexical cues generally, the disproportionate flagging of racial discrimination disclosures by algorithms is not simply driven by higher frequencies of these cues. Other linguistic features unique to racial discrimination disclosures may also play a significant role, particularly considering previous research highlighting the tendency of content moderation algorithms to overly react to identity references (e.g., markers of African American English) (10, 47).

Given that discussions about race that contain these specific markers may trigger algorithmic flagging, it is crucial to exercise caution against an overreliance on automated systems, especially given their limitations in contexts requiring nuanced discernment. Modern algorithms cannot yet distinguish such nuance when moderating race talk.

## Study 2a: Are Humans Better?

While algorithms provide an efficient and scalable approach to content moderation, many social media platforms actively engage users in flagging content that violates platform guidelines,

which leads to further review and, in some cases, removal from the platform (6–8, 48). In contrast to algorithms, humans have a greater capability to comprehend nuances and context, potentially making them better equipped to respond to racial discrimination disclosures. At the same time, racial discourse in the United States is often perceived as contentious or divisive, which may motivate humans to remove it (13). To explore this, we examined human content moderation responses (flagging done by real neighbors of the poster) to racial discrimination disclosures using metadata from our compiled datasets.

Although human flagging is typically infrequent on the studied platform, comprising only 2% of posted content overall, we found a significant surge in flagging rates for racial discrimination disclosures, reaching 36%. Even when compared with the control group of negative interpersonal experience disclosures, humans are almost three times more likely to flag racial discrimination disclosures (36.10% [33.27%, 39.12%]) compared to negative interpersonal experience disclosures (13.68% [11.60%, 16.06%]), $\chi^2(1) = 136.41$, $P < 0.0001$ (Fig. 1*B*). The stark contrast between the base rate of flagging and the flagging of racial discrimination disclosures challenges the assumption that humans would be more understanding than algorithms in their content moderation practices.

## Study 2b: What Predicts Human Flagging of Racial Discrimination Disclosures?

Unlike machines, humans are likely able to discern contextual nuances beyond affective lexical cues such as swear words. However, when confronted with discussions about racial discrimination—a topic often prompting discomfort and denial among individuals from dominant racial backgrounds—they may contend with social identity threat (13–16).

Research on social identity and intergroup relations suggests that people's judgments of others depend in large part on whether they are seen as a member of one's social group (e.g., racial group, political group), and that people are motivated to protect members of their ingroup (49–51). In the case of racial discrimination disclosures, if readers of a post perceive that members of their racial group are being accused of discrimination, it may instigate concern about being associated with discriminatory behavior ("Does this reflect negatively on my group or myself?") (16, 52). This identity threat could motivate individuals to suppress these discussions to protect their self-image and racial group image (15). In contrast, when readers perceive the poster as part of their ingroup, they may be more receptive to the poster's perspective (53). In the case of a local-based online community, posters and readers inherently share a social identity of living in the same neighborhood (54); when neighborhood ingroup status is made salient, readers may feel increased affinity for the poster and thus be less motivated to flag their disclosure (55, 56).

Finally, research has shown that people from majority groups may respond to identity threats by engaging in tone policing: critiquing the way that injustice is called out rather than engaging with the injustice itself (57, 58). Given that discussions about race often evoke discomfort, we predicted that human readers might seek a nonracial rationale for their discomfort by closely monitoring the negative emotional tone of racial discrimination disclosures, potentially using negativity as a justification for flagging content for removal (e.g., the poster sounds too angry) (59). For example, in Fig. 2*B*, both examples listed under "tone-policing" are from racial discrimination disclosures. These examples contain negative language (e.g., "It's horrible"), and were flagged by users. However, we find examples of negative interpersonal experience disclosures that contain similar language (e.g., "Horrible customer service") but were not flagged by users. We test whether this pattern persists throughout our dataset.

Altogether, we extracted three linguistic factors: psychological belonging, social identity threat, and tone policing (see Fig. 2*B* for examples). Lexicons were formalized leveraging the Fightin' Words method (60), a technique commonly used to identify statistically overrepresented words in a corpus of texts, relative to another corpus (see *Materials and Methods* for details). The

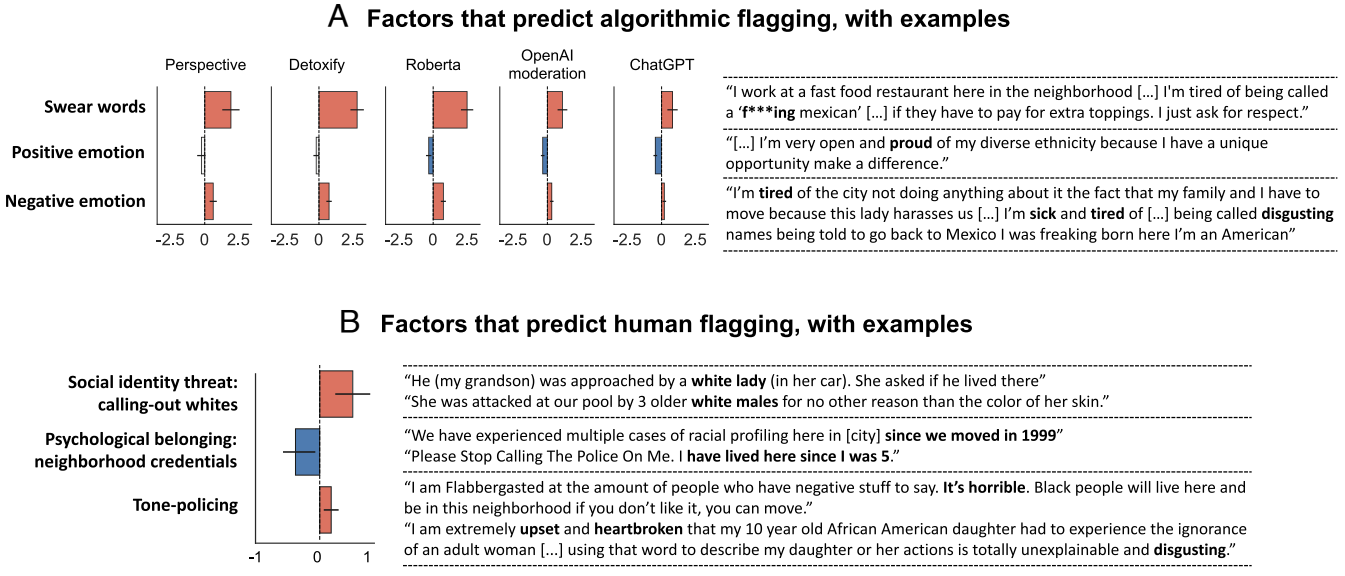


**Fig. 2.** (*A*) Lexical markers of affect predict algorithmic flagging (Study 1b). The effect estimate (on the x-axis) of the impact of the linguistic features (on the y-axis) on algorithmic flagging, across the five tested toxicity detection algorithms. Red indicates increased flagging of racial discrimination disclosures ($P < 0.05$); blue indicates reduced flagging of racial discrimination disclosures ($P < 0.05$). Error bars represent bootstrapped 95% CIs. (*B*) Psychological belonging and threat and tone policing predict human flagging (Study 2b). The effect estimate (on the x-axis) of the impact of the linguistic features (on the y-axis) on human flagging. Again, red indicates increased flagging of racial discrimination disclosures ($P < 0.05$), blue indicates reduced flagging of racial discrimination disclosures ($P < 0.05$). Error bars represent bootstrapped 95% CIs.

complete lexicons are outlined in *SI Appendix*, Table S1. Next, we fitted logistic regression models to measure the impact of these factors on flagging (*Materials and Methods*). The full regression modeling statistics are listed in *SI Appendix*, section 3.E.

**Psychological Belonging and Threat Drive Human Flagging.** The explicit mention of White racial identity, phrases such as "white man" or "white lady," was predictive of increased human flagging behavior ($\beta = 0.57$, $P < 0.001$) (Fig. 2*B*). Conversely, using language indicating one's belonging in the neighborhood, phrases such as "we've lived" or "since the 1980s," was linked to reduced human flagging ($\beta = -0.42$, $P < 0.05$). These analyses controlled for word length, emotion scores, and swear words (*Materials and Methods*).

Further robustness checks corroborated the influence of these particular psychological factors on human flagging rates (*SI Appendix*, section 3.G). Mentioning of Black racial identity (e.g., "black man," "dark lady," "blacks"), mentioning specific individuals (e.g., "this person," "this woman," "this man"), discussing the Black Lives Matter movement, or mentioning American credentials (e.g., "citizen," "veteran") did not significantly affect human flagging rates.

**Tone-Policing in Human Flagging.** The presence of negative emotion words was a predictor of human flagging for racial discrimination disclosures but not for negative interpersonal experience disclosures (*SI Appendix*, section 2.D). This distinction implies that humans are not reacting indiscriminately to negativity, but might be engaging in tone-policing regarding racial discrimination disclosures. Tone-policing occurs when individuals with privilege redirect the focus of the conversations from the content (e.g., about oppression) to the tone, language, or manner of discussion (61). The finding that negative emotion predicts flagging in racial discrimination disclosures but not negative interpersonal experience disclosures suggests that human raters might use emotional expression as a nonracial justification for taking down racial discrimination disclosures.

In sum, despite their capacity to understand linguistic nuances, humans flag racial discrimination disclosures for removal. These findings highlight the need to better understand and address the psychological and social influences shaping human behaviors in the online content moderation space, particularly in conversations related to identity.

## Study 3: Does Racial Suppression Harm black Onlookers?

Across Studies 1a to 2b, we found that both prominent algorithms and humans alike disparately suppress discussions about racial discrimination. In Study 3, we examine whether seeing a racial discrimination disclosure being reported for removal impacts other users' sense of belonging and connection to both their physical and digital communities. Here, we focus on Black Americans, as discrimination experiences from Black Americans were most prevalent in our labeled dataset (*SI Appendix*, section 1.D), which is also reflective of a broader pattern of discrimination in the United States (62).

We recruited 338 Black Americans through an online recruitment platform and presented them with an online neighborhood page which simulated a real social media feed, displaying multiple posts concurrently. Within this feed, participants were randomly assigned to view either a racial discrimination disclosure or a negative experience disclosure. These posts were either flagged for removal by a neighbor or not flagged, constituting a 2 (Content: racial discrimination disclosure vs. negative experience disclosure) x 2 (Flagged for removal: flagged vs. unflagged) experimental design. We simulated the experience of learning about flagging by adding a warning label to the post of interest, indicating that it had been flagged for removal by another user and that it would shortly be removed from the platform. While this design choice does not reflect all social media feeds [although some have used similar labels (63)], it allows us to manipulate the knowledge of flagging, and thus measure its impact (see *SI Appendix*, section 4.D for possible sources of moderation awareness).

After viewing the neighborhood social media page (including the focal post), participants were asked how they would feel about the neighborhood, its residents, and the platform itself, as well as about the posted content (*SI Appendix*, section 4.A). We hypothesized that viewing a flagged racial discrimination disclosure, relative to other conditions, would result in more negative perceptions of the neighborhood, neighbors, and the platform. We also hypothesized that Black participants would value discrimination disclosures more than negative interpersonal experience disclosures and would be more upset about their removal. Code and data for this study are available at https://osf.io/f3eqt/.

**Connection to the Neighborhood.** After viewing the neighborhood page, participants were asked about their connection to the neighborhood using five items (e.g., "Given an opportunity, I would like to live in this neighborhood"; 1 = Strongly disagree, 7 = Strongly agree; $\alpha = 0.89$). First, participants who viewed the racial discrimination disclosure post indicated significantly weaker feelings of neighborhood connection compared to those exposed to the negative experience disclosure, possibly as the racial discrimination disclosure indicated explicit evidence of racism present in the neighborhood ($F(1, 328) = 55.39$, $P < 0.001$, $\eta_p^2 = 0.14$). There was not a significant main effect of seeing a post flagged on feelings of neighborhood connection ($F(1, 328) = 2.63$, $P = 0.11$, $\eta_p^2 < 0.01$). However, as hypothesized, an interaction effect emerged [$F(1, 328) = 4.85$, $P < 0.05$, $\eta_p^2 = 0.01$)]. Participants exposed to the racial discrimination disclosure post flagged for removal reported reduced feelings of neighborhood connection compared to those exposed to the same disclosure without any flagging indication (flagged: M = 3.25, SD = 1.37 and unflagged: M = 3.77, SD = 1.28). Conversely, the flagging status of the negative experience disclosure did not influence feelings of neighborhood connection (flagged: M = 4.56, SD = 1.15; unflagged: M = 4.48, SD = 1.09) (Fig. 3).

**Perception of the Platform.** When asked how likely they would be to use this social media platform (1 = Not at all inclined, 5 = Extremely inclined), participants exposed to the racial discrimination disclosure exhibited a marginal decrease in their inclination to use the platform ($F(1, 328) = 3.83$, $P = 0.05$, $\eta_p^2 = 0.01$). The presence of a flagged post did not significantly influence participants' inclination to use the platform ($F(1, 328) = 2.36$, $P = 0.13$, $\eta_p^2 < 0.01$). There was a marginal interaction between content type and flagging status ($F(1, 328) = 3.76$, $P = 0.05$, $\eta_p^2 = 0.01$): Among participants exposed to the racial discrimination disclosure, those witnessing its flagging experienced a diminished inclination to use the platform (M = 2.36, SD = 1.33) compared to those who did not see any indication of flagging (M = 2.84, SD = 1.33). Conversely,

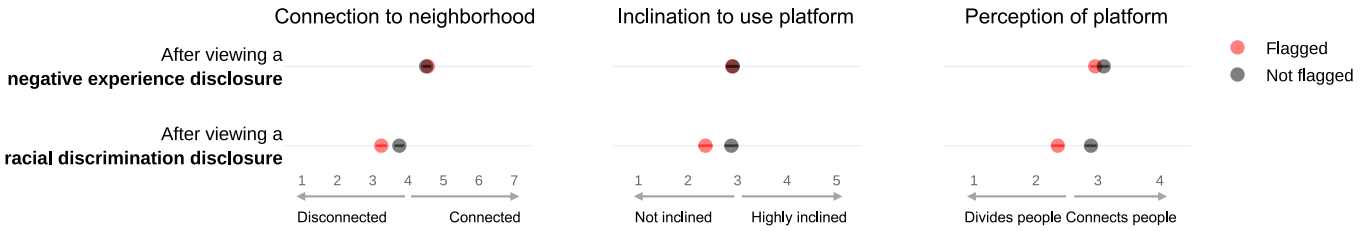


**Fig. 3.** Impacts on relationship to neighborhood and relationship to platform. Participants who saw the racial discrimination disclosure being flagged for removal felt lower connection to the neighborhood, expressed less willingness to use the platform, and saw the platform as a more divisive place, compared to participants who did not see a flagged racial discrimination disclosure. Flagging did not influence these variables in the negative experience disclosure conditions. Error bars (darker lines within circles) indicate SE of the mean.

among those who viewed the negative experience disclosure, no significant difference in inclination to use the platform emerged based on flagging status (flagged: M = 2.90, SD = 1.22; unflagged: M = 2.85, SD = 1.15) (Fig. 3).

Participants were also asked to what extent they believed the platform had the potential to unite or divide people (1 = Very much divide people, 4 = Very much bring people together). Those who viewed the racial discrimination post expressed a reduced belief in the platform's ability to promote social cohesion compared to those who encountered the negative experience disclosure ($F(1, 328) = 18.08$, $P < 0.001$, $\eta_p^2 = 0.05$). Participants perceived the platform as less capable of social cohesion after seeing a flagged post compared to instances where there was no indication of a flag ($F(1, 328) = 11.25$, $P < 0.001$, $\eta_p^2 = 0.03$). As hypothesized, there was a significant interaction between content type and flagging status ($F(1, 328) = 5.02$, $P < 0.05$, $\eta_p^2 = 0.03$). Specifically, participants exposed to the flagged racial discrimination disclosure perceived the platform as significantly less capable of bringing people together (M = 2.37, SD = 0.91) compared to participants who did not see any indication of flagging (M = 2.88, SD = 0.88). However, among participants exposed to the negative experience disclosure, their perception of the platform's potential for fostering social cohesion remained consistent regardless of flagging status (flagged: M = 2.96, SD = 0.73; unflagged: M = 3.10, SD = 0.71) (Fig. 3).

**Seeing Racial Discrimination Disclosures as Valuable.** We next asked participants how appropriate either the racial discrimination disclosure or the negative experience disclosure was to share on a neighborhood platform. Participants rated the racial discrimination disclosure as more appropriate to share on a neighborhood platform in comparison to the negative experience disclosure, ($F(1, 328) = 51.54$, $P < 0.0001$, $\eta_p^2 = 0.14$); there was no significant effect of flagging status, ($F(1, 328) = 0.12$, $P = 0.73$, $\eta_p^2 < 0.001$). There was also no significant interaction between content type and flagging status, ($F(1, 328) = 0.36$, $P = 0.55$, $\eta_p^2 < 0.01$).

Participants also rated how good and bad (reverse-coded) the post was for the neighborhood (1 = Strongly disagree, 7 = Strongly agree, $r = 0.72$). The racial discrimination post was perceived as significantly more beneficial for the neighborhood compared to the negative experience disclosure ($F(1, 328) = 19.66$, $P < 0.001$, $\eta_p^2 = 0.05$). Participants rated flagged posts as marginally better for the neighborhood than unflagged posts ($F(1, 328) = 3.64$, $P = 0.06$, $\eta_p^2 = 0.01$). There was no significant interaction between content type and flagging status ($F(1, 328) = 0.46$, $P = 0.50$, $\eta_p^2 < 0.01$).

Participants were also asked about how upset they were that this content was removed (or in the unflagged condition, how upset they would be). Participants expressed higher emotional distress when a racial discrimination disclosure was removed, compared to the removal of a negative experience disclosure ($F(1, 327) = 125.61$, $P < 0.0001$, $\eta_p^2 = 0.28$). Flagging status did not influence emotional distress, and there was also no significant interaction between content type and flagging status.

**Educational Potential of Racial Discrimination Disclosures.** Participants endorsed the educational capacity of discrimination disclosures. Three items measured perceived discrimination awareness in the neighborhood (e.g., "People in this neighborhood think it's important to discuss discrimination experiences in the community"; 1 = Strongly disagree, 7 = Strongly agree, $\alpha = 0.76$). While there was no difference by the type of content ($F(1, 328) = 0.12$, $P = 0.73$, $\eta_p^2 < 0.01$), participants who saw content being flagged perceived lower discrimination awareness in the neighborhood ($F(1, 334) = 42.82$, $P < 0.0001$, $\eta_p^2 = 0.12$). There was a significant interaction between content type and flagging status ($F(1, 328) = 26.75$, $P < 0.0001$, $\eta_p^2 = 0.08$). Participants who saw the racial discrimination disclosure without any indication of flagging recognized that people in the neighborhood think it is important to discuss discrimination experiences in the community (M = 4.82, SD = 0.94). However, when there was an indication of flagging of the racial discrimination disclosure, participants indicated that this neighborhood has lower discrimination awareness (M = 3.21, SD = 1.69). Meanwhile, among participants exposed to the negative experience disclosure, their perception of discrimination awareness remained consistent regardless of flagging status (flagged: M = 3.75, SD = 1.25; unflagged: M = 3.94, SD = 0.95).

In summary, our findings reveal that when posts addressing experiences of discrimination are subject to suppression, it triggers a sense of disconnection among Black observers in the community. Seeing a racial discrimination disclosure being reported for removal also diminishes their inclination to use the platform, and leads them to see the platform as more divisive. It is worth noting that this negative reaction is not directed at the content of the discrimination disclosures itself. On the contrary, participants acknowledge the valuable role these disclosures play in raising awareness within the community. This study provides evidence that racial discrimination disclosures can serve an important purpose. Moreover, suppression of such disclosures can erode the sense of belonging in the community, both online and offline.

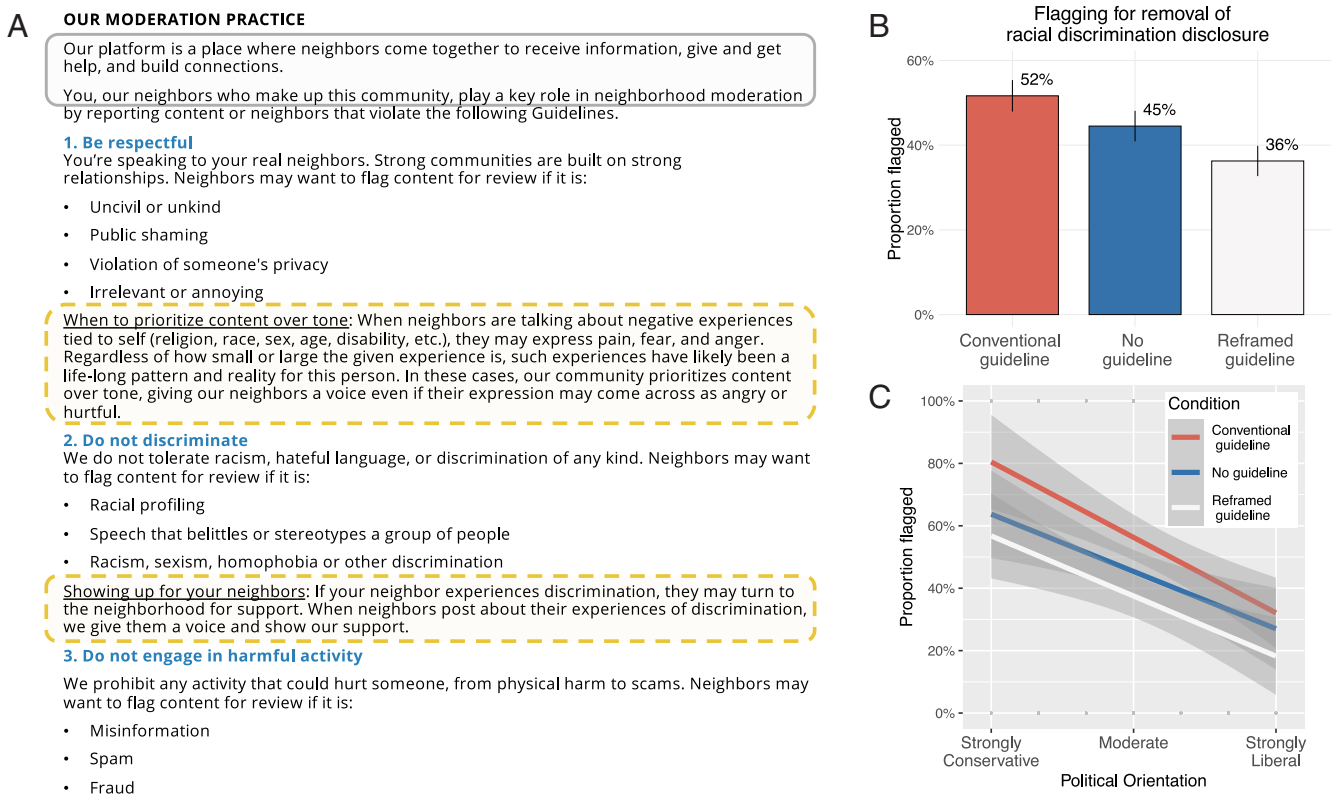


**Fig. 4.** (*A*) Moderation Guidelines used in Study 4. Participants in the no-guideline condition only saw content within the gray solid line. Participants in the conventional guideline and reframed guideline conditions saw content within the gray solid line and the enumerated guidelines. The reframed guideline condition additionally saw content within the yellow dotted lines. (*B*) Flagging rates of discrimination disclosures by condition. Error bars mark SE of the mean. (*C*) Flagging rates of discrimination disclosures by condition and political orientation.

## Study 4: Can We Reduce Suppression of Racial Discrimination Disclosures?

Studies 1 to 3 uncover suppression of discrimination disclosures and its downstream negative impact on observers. In Study 4, we investigate how conventional content moderation guidelines may inadvertently encourage human flagging of racial discrimination disclosures, and whether reframing these guidelines can reduce flagging behavior. Conventional guidelines provide instruction on what types of content should be flagged, potentially promoting punitiveness among guideline readers. In our reframed guidelines, we clarify that racial discrimination disclosures are not inherently in violation of guidelines, thus encouraging the use of caution. Drawing insights from Study 2b, we draw attention to the dangers of tone policing, and highlight collective identity to mitigate social identity threat (see Fig. 4*A* for the intervention language).

Given that referencing a shared identity decreased the likelihood of flagging racial discrimination disclosures (Study 2), we emphasized the shared overarching identity of "neighbor" in the reframed guidelines. Phrases such as "your neighbor" and "our community" were employed to redirect attention to this collective identity when referring to the target of discrimination and the reader. We theorized that shifting the focus to a shared ingroup could foster increased understanding of and empathy for the poster (64, 65).

Study 2 also found that humans engaged in tone-policing of racial discrimination disclosures by flagging those with a more negative tone. To explicitly address this, we outlined criteria for when to prioritize content over tone in our reframed guidelines. Additionally, we incorporated the perspective of a target of discrimination into the reframed guidelines, as gaining perspective can provide a more accurate view of another's experience (66), and has been shown to increase empathy for a target and their entire group (67, 68). The reframed guidelines highlight the fact that any single experience of discrimination could be part of a recognizable pattern for the target, rather than an isolated incident. We theorized that highlighting this possibility would offer valuable context for the negative emotional tone often found in racial discrimination disclosures.

Guided by results from Study 2, which revealed elevated rates of flagging in instances where White identity is mentioned, we direct this initial effort toward reducing flagging among White American participants. In an online experiment similar to Study 3, we presented 555 White American participants with an online neighborhood page. Participants were randomly assigned to view either no moderation guidelines, conventional guidelines, or our reframed guidelines. Next, participants were asked whether they would flag a series of posts, one of which was a racial discrimination disclosure (the rest were filler posts). To test for condition differences in flagging of racial discrimination disclosures, we fit a logistic regression model to test for the binary decision to flag the post for removal. Code, data, and preregistration for this study are available at https://osf.io/f3eqt/.

**Does Reframing Guidelines Reduce Flagging?** As hypothesized, having participants read the reframed guidelines led to fewer instances of flagging discrimination disclosures (36.26%) compared to those who read and followed the conventional platform guidelines (51.65%; $\beta = -0.79$ [−1.23, −0.35], SE = 0.22,

z = −3.51, $P < 0.001$) (Fig. 4*B*). Participants in the no-guideline condition were also less likely to flag a discrimination disclosure (44.50%) than those who saw conventional platform guidelines ($\beta = -0.44$ [−0.87, −0.01], SE = 0.22, z = −2.01, $P < 0.05$). Deviating from our preregistered hypothesis, the flagging rate of participants in the reframed guideline condition was not significantly lower than that of participants in the no-guideline condition ($\beta = 0.35$ [−0.08, 0.78], SE = 0.22, z = 1.60, $P = 0.11$).

As political polarization has increased in recent years (69), and has been facilitated in part by social media (70), it is important to consider whether efforts to shift behavior may sow further division. Thus, we tested for an interaction between condition and political ideology on flagging, assessing whether our intervention had varying effects by political group. While there is a main effect of political orientation, such that conservative participants were more likely to flag the racial discrimination disclosure than liberal participants, $\beta = -0.34$ [−0.520, −0.178], SE = 0.09, z = −3.95, $P < 0.001$ [which is consistent with prior work (47)], we found no significant interaction between condition and political orientation (Fig. 4*C*). This is initial evidence that our intervention may provide a strategy for reducing the suppression of racial discrimination disclosures among users across the political spectrum, though more work is needed to further investigate the robustness of this effect as well as whether and how political motivation relates to flagging behavior.

The current study replicated the human flagging observed in Study 2. We found, using a broad online participant pool, that 52% of White Americans presented with conventional platform guidelines flagged the racial discrimination disclosure for removal. Our intervention significantly reduced the flagging of racial discrimination disclosures to 36%. We consider this intervention a conservative test, as it retained all of the language in the conventional guidelines that encourages flagging across different content categories. The finding demonstrates that even within an environment where proactive flagging is primed and encouraged as a hallmark of responsible user conduct, implementing minor adjustments to shift perspective and encourage empathy can alleviate the suppression of racial discrimination disclosures.

Relatedly, our findings suggest the emphasis on flagging as a moderation tool might inadvertently make even unwarranted flagging behavior seem acceptable (71). In both the conventional and reframed guideline conditions, participants were encouraged to see flagging as a key behavior of responsible user conduct ("You, ..., play a key role in neighborhood moderation by reporting content or neighbors that violate the following guidelines."), and were given specific guidelines for flagging ("neighbors may want to flag content for review if it is: uncivil [..]"). In contrast, those in the no guideline condition received neither a mention of flagging nor specific guidelines before reviewing posts. These two components might unintentionally create an allowance for users to flag content even when it does not violate guidelines. For instance, users might opt to flag content that feels questionable or uncomfortable, attributing it to proactive moderation. This might explain the higher flagging rates of racial discrimination disclosures in the conventional guideline condition compared to the no-guideline condition, and the lack of significant difference between our intervention and the no-guideline condition (36.26% [29.28%, 43.24%] vs. 44.50% [37.45%, 51.55%], respectively). This would also suggest that the context added to the specific guidelines in the intervention condition ("When to prioritize content over tone [...]," "Showing up for your neighbors [...]") was effective enough to counteract the encouragement of flagging specifically for racial discrimination disclosures. However, future work should disentangle the effects of encouraging flagging behavior and spelling out specific guidelines on rates of racial discrimination disclosure flagging. The nonsignificant difference between the intervention and no-guideline conditions may also be due to insufficient power, which future work may investigate.

At the same time, we caution that having no guidelines or tools for moderation may prompt genuinely harmful content to proliferate unhindered. Indeed, previous research found decreased civility in comments when no guidelines are provided (72). More work is needed to disentangle the different elements of moderation and their impact on various types of content (both harmful and not harmful). Future research should also explore alternative mechanisms to reinforce platform norms and rules, encouraging more prosocial and inclusive online behaviors while still deterring harmful behavior (73). Meanwhile, given that moderation guidelines remain a predominant tool for platforms, our intervention demonstrates that even small changes to reframe these guidelines can reduce suppression of racial discrimination disclosures.

## Discussion

Our research shows that users who share experiences of racial discrimination, crucial for fostering empathy, social support, and meaningful conversations, are disproportionately silenced by both modern algorithms and humans. Five state-of-the-art and publicly available off-the-shelf algorithms employed by major companies all suppressed racial discrimination disclosures. While humans also displayed bias against racial discrimination disclosures, providing them with psychologically informed reframing mitigated the influence of such bias.

The present work drew upon multiple sources and interdisciplinary methodologies. Initially, it involved compiling and labeling posts written by real neighbors, enabling a field study on how modern algorithms treat these posts (Study 1a), alongside behavioral data analysis from platform users (Study 2a). We used computational linguistic tools to dissect predictive factors that informed the intervention (Study 1b and 2b). In an online survey, we documented impact on of Black American participants (Study 3), and we experimentally tested an intervention with White American participants (Study 4). As far as we know, these are the first studies to systematically examine content moderation behavior toward racial discrimination disclosures, identify its far-reaching repercussions, and take initial steps toward potential solutions.

Our investigation has delineated two distinct categories of harm emanating from biased moderation of racial discrimination disclosures. The first, direct harm (Study 1a and 2a), is rooted in the suppression itself. Algorithmic and human flagging leads to content removal or down-ranking, effectively silencing the voices of individuals addressing racial discrimination, preventing them from receiving the support they need and thwarting the opportunity to initiate constructive discussions. The second, indirect harm (as examined in Study 3), is a harm that extends beyond the direct targets of silencing to onlookers. This form of harm manifests in a sense of disconnection with the neighborhood and a reduced inclination to use the platform, despite its potential benefits. Consequently, suppression of racial discrimination disclosures can lead individuals, even those who have not been directly silenced themselves, to experience isolation from both community members and platform resources.

Our research presents a potential mitigation to the suppression of racial discrimination disclosures—a psychologically informed intervention that reframes conventional moderation guidelines leveraged by online platforms to reduce human flagging. Given our findings that humans flag due in part to social identity threat, moderation guidelines can be framed in ways that ameliorate that threat and promote perspective taking, shared identity, and prosocial norms.

**Future Directions.** The effectiveness of the guideline reframing intervention among human participants naturally gives rise to the question of how toxicity classifiers fare in comparison (see *SI Appendix*, section 2.C for discussion on testing debiased models and GPT-4). One might speculate, for instance, whether correcting biases in machines proves more tractable than addressing deep-seated human biases. For example, one could provide the algorithms with better-informed labels to serve as the ground truth during training (74). Although we lack direct access to the training data of toxicity algorithms, preventing us from directly testing this hypothesis, we suspect that alongside the broader influence of human biases often present in large datasets (75), the process of curating data for training toxicity algorithms can introduce additional layers of complexity. For example, toxicity datasets are often curated based on targeted keyword searches due to the rarity of toxic content relative to nontoxic content (76). These heuristic approaches in data sampling can cause toxic speech classifiers to learn spurious lexical correlations while lacking comprehensive contextual understanding, similar to our findings in Study 1b (77). Future work could explore ways to provide algorithms with better social and contextual understanding, including explicitly embedding societal values (78).

While the current research recruited online samples of Black Americans (Study 3) and White Americans (Study 4), our examination of the racial discrimination disclosure dataset revealed active participation in discussions about racial discrimination and inequity across many different racial groups (*SI Appendix*, section 1.D). To develop a more comprehensive understanding of how online suppression affects different demographics and their interactions in digital spaces, future studies should extend our approach to include a more extensive spectrum of voices and experiences, both within and outside of the United States. Moreover, while we have identified specific linguistic predictors of flagging behavior, numerous other factors likely influence such behavior. For instance, platform structure (e.g., organized around interest, personal relationship, or career) may result in different patterns, reasons, and reactions to suppression. Similarly, broader platform norms and practices regarding challenging yet potentially constructive discussions, particularly surrounding diversity and equity, may significantly influence the perception and treatment of such conversations (79). Recent work also suggests that an individual's perceptions of any particular post are shaped by their attitudes toward moderation practices (80) and broader societal issues, such as racial discrimination (47). Our analysis represents just one facet of this complex puzzle.

We show that conversations about discrimination are often prevented through flagging. But when these conversations are able to unfold, how do people respond? Future research should explore prevalent responses to racial discrimination disclosures and their impact on community members, shedding light on the dynamics of race-related conversations and their potential to drive or hinder societal change. For example, researchers should investigate the extent to which responses make the discloser feel understood, how emerging technology may support such conversations to be more empathetic, and whether this improves conversational outcomes (81).

Finally, while our guideline reframing intervention shows initial promise, future work should continue to investigate the predictors of suppression and other potential mechanisms for further reducing it. For example, future work could isolate the components of the current intervention to determine their respective roles in flagging reduction. Future work could also test this intervention on a major social media platform in order to assess its viability in a field setting over time, as well as whether the intervention has positive spillover effects on race and identity-related conversations in general, beyond racial discrimination disclosures. Finally, there is potential for similar psychologically relevant strategies and platform design interventions to be tested in promoting prosocial behaviors online (see ref. 82 for an example).

**Conclusion.** Fostering constructive conversations about race remains an open problem. Our work sheds light on the ways these conversations get suppressed by a wide range of algorithms used by social media platforms, and how conventional moderation practices have the potential to both exacerbate and mitigate suppression.

In a world where online communication is increasingly integral to our lives, it is imperative that we address matters of content moderation and suppression (83), especially when it comes to issues of racial discrimination. Our research endeavors to pave a path forward, one that ensures the positive potential of social media is realized without stifling the voices of those who have historically been marginalized.

Our work highlights the pressing need to rethink content moderation guidelines, algorithms that enforce them, and human moderation practices such that they closely reflect societal values, such as the values of inclusivity and equity, in neighborhoods and beyond.

## Materials and Methods

### Data Identification.

***Data sharing.*** We utilize data from a social networking platform used in neighborhoods across every state in the United States. This platform is a social network for people in the neighborhood to share information, help out one another, and connect. The platform includes representation across different demographic groups, including 8% of African Americans, 7% of Asian Americans, 14% of Hispanic/Latino Americans, and 80% of White Americans.

Stanford SPARQ is a behavioral science "do tank" that builds research-driven collaborations with private and public sector organizations. Through the center and Stanford University, our research team has established a data usage and research collaboration agreement with the platform. The agreement includes terms for research independence, publication rights, and data integrity. It also ensures a secure data pipeline to transfer, store, and analyze data, as well as anonymize and safeguard confidential data.

***Data sampling.*** The initial dataset is a random sample of around 30 million posts that had been posted across the United States for the entire year of 2020, as well as comments on the posts, and replies to the posted comments. Given the large size of the initial random dataset, we utilized additional filters to find posts that shared personal experiences. Thus, from the original dataset, we further filtered down to posts containing a personal entity keyword (e.g., "my son") and an experience keyword (e.g., "encounter"). All keywords are listed in *SI Appendix*, section 1.A.

To further identify race-related personal experience disclosures, we additionally implemented two-step filters. First, we filtered for conversations that contained a race-related keyword (e.g., "Black," "racism"). Second, finding that this filter was still relatively broad, we used a bootstrap-like technique to compile a set of higher-precision race-related keywords, and

sampled for posts containing one of these higher-precision race-related keywords. The high-precision keywords and the procedure are listed in *SI Appendix*, section 1.B.

***Data annotation.*** Ten trained coders–graduate students in psychology, linguistics, and computer science, as well as two undergraduate research assistants–annotated the samples (Method 1.B). Each coder annotated an overlapping subset with another coder. Pairs reached moderate to strong agreement: interrater reliability ranged from $\kappa = 0.70$ to $\kappa = 0.92$ for racial discrimination disclosures and $\kappa = 0.75$ to $\kappa = 0.89$ for negative interpersonal experience disclosures. Disagreements were resolved by a third coder.

**Flagging of Racial Discrimination Disclosures.**

***Algorithms.*** To calculate algorithmic toxicity scores, we utilized a set of commonly used preexisting models: 1) The Perspective automated programming interface, 2) The Detoxify, and 3) The Roberta toxicity classifier. These models employ machine learning algorithms to identify text that exhibits toxic behavior. They have been trained or fine-tuned on supervised toxicity classification tasks. In addition, we also employed 4) A ChatGPT-based automated programming interface (GPT-4 model) and 5) OpenAI's Moderation automated programming interface[†] to annotate the studied posts. Toxicity scores and annotations were extracted in Dec 2023. Perspective and Detoxify give only a score (standard threshold of 0.5 was used), while ChatGPT, OpenAI moderation API, and Roberta give a binary flagged label. See *SI Appendix*, section 2.A for details on how ChatGPT was prompted for toxicity classification, including prompt text and prompting variants. For analysis, we process the text of each post using these tools, which allowed us to compute a toxicity score for each post. In cases where the interfaces imposed restrictions on input length, we ensured that the analysis was performed using the maximum allowable number of tokens for each post.

***Humans.*** We compared the flagging percentage difference between two types of posts: racial discrimination disclosures ($N = 1{,}025$) and negative interpersonal experience disclosures ($N = 1{,}009$). Flagging may be done either by regular users or volunteer moderators responsible for regulating content that violates the platform's guidelines. Users and moderators on this specific platform are neighbors within the same geographic area where the posts are made.

**Linguistic Analyses of Factors Influencing Flagging.**

***Feature extraction.*** To analyze factors influencing flagging, we utilized the Fightin' Words method (60), a technique commonly used to identify statistically overrepresented words in a corpus of texts, relative to another corpus. Insights derived from the Fightin' Words method guided the formalization of lexicons for Neighborhood credentials and Calling out of White identity (*SI Appendix*, Table S1). This process of lexicon induction is a standard approach in the field (84–86). The specific implementation of the linguistic features we selected to test our hypotheses is outlined in *SI Appendix*, Table S1. To estimate positive and negative emotion for a given content, we use Vader, an emotion-scoring algorithm specialized for social media posts (36). Swearwords were defined as the swearwords category in LIWC lexicons (37). Other features are implemented based on lexicons.

***Modeling approach for factors influencing flagging.*** All features were implemented using the Python programming language. The linguistic features were coded in binary form, indicating the presence or absence of a token. However, the word count, positive emotion score, and negative emotion score were continuous variables that were standardized to have a mean of zero and a SD of one for ease of comparison. We fitted a logistic regression model, using standard Python libraries numpy and statsmodels (87, 88). One data point corresponds to a racial discrimination disclosure post ($N = 1{,}025$). Our model included various control features: post length, emotion scores, and the presence of swear words (*SI Appendix*, section 3.E). We used a threshold of 0.5 to determine flagging as the algorithms produced a score within the range of 0 to 1.

**Impact.**

***Participants.*** We recruited a sample of 338 Black/African Americans from CloudResearch. We limited our sampling to individuals aged 25 and above to target those who have a greater probability of making housing decisions on their own. There was no attrition. The final sample comprised 338 participants [68% women, 32% men; mean age = 39.84 (SD = 10.63)]. All procedures for this and subsequent experiments were approved by the ethics board at Stanford University (Protocol No. 57650).

***Procedure.*** This study took the form of a 2 (Content: racial discrimination disclosure vs. negative experience disclosure) x 2 (Flagged for removal: flagged vs. unflagged) between-subjects design. Following their consent to participate, participants were given basic information about a neighborhood that was majority White. Then they were presented with an online neighborhood page and asked to imagine what this neighborhood and the neighbors would be like. The design of the page mimicked a conventional social media interface, with five posts related to the neighborhood displayed in a feed. Among these posts, four were filler posts, while one post was the focal point of our study. The post of interest was consistently presented as the third post. Participants were randomly assigned to read either a post disclosing a racial discrimination experience or a post recounting a negative interpersonal experience that was not related to race (see *SI Appendix* for stimuli selection details). Our stimuli set consisted of four pairs of racial discrimination disclosures and negative interpersonal experience disclosures that were matched on word length and emotion. Additionally, participants were randomly assigned to either a flagged or unflagged condition. Those in the flagged condition were informed that the post of interest would be removed as a result of flagging, while the unflagged condition lacked this notification.

After viewing the neighborhood group page, participants rated their connection to the neighborhood, the neighborhood's attitude toward discrimination, their own thoughts about the platform, as well as their perception of the racial discrimination disclosure as appropriate. Finally, participants answered basic demographic questions and were paid for their time. Further information, including all items included in study 3, can be found in *SI Appendix*, section 4.A.

***Analysis.*** We ran an ANOVA for each outcome, with Content (racial discrimination disclosure vs. negative experience disclosure) and Flag (flagged vs. unflagged) as predictors. Simple effects analyses were corrected with Bonferroni multiple testing corrections. Given the gender imbalance of our sample (68% women, 32% men), we ran an additional model for each outcome with gender as a moderator, as well as a model with political orientation as a moderator. There was no main effect of political orientation, nor any interactions. We found a Content × Flag × Gender interaction for connection to neighborhood, such that women feel less connected to the neighborhood when seeing a racial discrimination disclosure flagged, while men are less impacted by flagging status of a racial discrimination disclosure. We did not find this interaction in any other outcomes (*SI Appendix*, section 4.B). Code and data for this study are available at https://osf.io/f3eqt/.

**Intervention.**

***Participants.*** Based on a power calculation to detect an effect size of $f = 0.14$ at 85% power, we needed a total sample of 561 participants. We recruited 600 White Americans from CloudResearch, to account for potential exclusions. Similar to Study 3, we limited our sampling to individuals over 25 years old. One participant did not finish the survey. 44 participants who did not identify as White/European American or identified as multiracial or biracial were excluded from analysis, as specified in the preregistration. The final sample included 555 participants [71% women, 28% men, 1% nonbinary; mean age = 42.36 (SD = 12.37)].

***Procedure.*** This study included three conditions: no-guideline, conventional guideline, and reframed guideline. The primary dependent variable was whether participants flagged the racial discrimination disclosure post for removal. Participants were given basic information about a neighborhood that was majority White, as in the previous study. They then read the community guidelines presented in a format resembling a conventional social media interface. In the no-guideline condition, participants read a general introduction to the neighborhood. Those in the conventional guideline condition read guidelines promoting the principles to be respectful, to not discriminate, and to not engage in harmful activity. Participants in the reframed guideline condition received additional context aimed at reframing discrimination disclosures, based on insights from Study 2. Subsequently, participants were randomly shown a total of five posts, one at a time, and were asked for their decision to flag the post

[†] https://platform.openai.com/docs/guides/moderation/overview.

for removal. Afterward, participants were asked how they felt about the post, the poster, and the guidelines. Finally, participants answered basic demographic questions and were paid for their time.

***Analysis.*** We fitted a logistic regression model for each outcome, with the condition as the predictor and political orientation as a covariate. We also fit a regression model including gender as a covariate. We found no moderation by gender for flagging rates of discrimination disclosures (*SI Appendix*, section 5.B). Code, data, and preregistration for this study are available at https://osf.io/f3eqt/.

**Data, Materials, and Software Availability.** Some study data available (Due to the inclusion of sensitive and personally identifiable information in our datasets of social media posts, we are unable to make these datasets public. Code used to analyze data in studies 1 to 4 will be made available, as well as survey and experiment data and materials from studies 3 and 4 at https://osf.io/f3eqt/ (89)).

**ACKNOWLEDGMENTS.** This research was supported by the Russell Sage Foundation (Grant#: 2210-39869), with additional support from the Stanford Institute for Human-Centered Artificial Intelligence, Stanford University's Ethics, Society, and Technology Hub, the Open Phil AI Fellowship, and the Swiss NSF (Grant#: P500PT-211127). We thank our industry partner for their collaboration and data sharing. We also thank the reviewers for their feedback. The funding, collaborations, and research referenced in this publication were administered and supported by Stanford SPARQ, a center that builds research-driven partnerships with industry leaders and changemakers to combat bias, reduce disparities, and drive culture change. We thank the researchers at Stanford SPARQ for their assistance and feedback; Vedika Kanchan, Serena Lee, Kate Cressey, and Daphne Muhammad for research assistance; Justine Zhang for developing data filtering methods; and the Race and Social Inequality Lab, the Hazelnuts Lab, and the Dweck–Walton Lab for feedback.

Author affiliations: [a]Department of Psychology, Stanford University, Stanford, CA 94305; [b]Stanford SPARQ, Department of Psychology, Stanford University, Stanford, CA 94305; [c]Department of Computer Science, Stanford University, Stanford, CA 94305; [d]Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH 03755; [e]Department of Linguistics, Stanford University, Stanford, CA 94305; and [f]Graduate School of Business, Stanford University, Stanford, CA 94305

Author contributions: C.L., K.G., P.R.K., M.H., E.D., K.L.S., X.Z., M.G.H., H.R.M., D.J., and J.L.E. conceptualized the study; C.L., P.R.K., M.H., E.D., K.L.S., N.S., D.T., X.Z., M.G.H., H.R.M., D.J., and J.L.E. conducted data acquisition; C.L., P.R.K., M.H., E.D., K.L.S., N.S., D.T., and D.J. performed data curation; C.L., K.G., and P.R.K. developed the software; C.L., K.G., P.R.K., M.H., X.Z., M.G.H., H.R.M., D.J., J.L.E. (studies 1a-2b), and C.L., M.H., H.R.M., J.L.E. (studies 3 and 4) designed the methodology; C.L., K.G., P.R.K., M.H. (studies 1a-2b), and C.L., M.H. (studies 3 and 4) conducted the investigation; C.L. (studies 3 and 4), and K.G. (studies 1a-2b) performed the formal analysis; C.L., K.G., P.R.K., M.H., D.J., and J.L.E. wrote the original draft; C.L., K.G., M.H., X.Z., M.G.H., H.R.M., D.J., and J.L.E. reviewed, edited, and wrote the paper; and H.R.M., D.J., and J.L.E. supervised the study.

Reviewers: C.A.B., Duke University; and M.S., Carnegie Mellon University.

1. B. Auxier, M. Anderson, Social media use in 2021. Pew Research Center. https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/. Accessed 4 December 2023.
2. M. Anderson, S. Toor , How social media users have discussed sexual harassment since #MeToo went viral. Pew Research Center. https://www.pewresearch.org/short-reads/2018/10/11/how-social-media-users-have-discussed-sexual-harassment-since-metoo-went-viral/. Accessed 4 December 2023.
3. M. Duggan, 1 in 4 black Americans have faced online harassment because of their race or ethnicity. Pew Research Center. https://www.pewresearch.org/short-reads/2017/07/25/1-in-4-black-americans-have-faced-online-harassment-because-of-their-race-or-ethnicity/. Accessed 4 December 2023.
4. E. Vogels, The state of online harassment. Pew Research Center. https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/. Accessed 4 December 2023.
5. Meta, Facebook (2023). https://about.meta.com/technologies/facebook-app/. Accessed 4 December 2023.
6. Twitter, Help center–report violations (2023). https://help.twitter.com/en/rules-and-policies/x-report-violation. Accessed 4 December 2023.
7. TikTok, Report a video (2023). https://support.tiktok.com/en/safety-hc/report-a-problem/report-a-video. Accessed 4 December 2023.
8. Reddit, Reddit content policy (2023). https://www.redditinc.com/policies/content-policy. Accessed 4 December 2023.
9. T. Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media* (Yale University Press, 2018).
10. M. Sap, D. Card, S. Gabriel, Y. Choi, N. A. Smith, "The risk of racial bias in hate speech detection" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, L. Márquez, Eds., (Association for Computational Linguistics, 2019), pp. 1668–1678.
11. P. Fortuna, M. Domínguez, L. Wanner, Z. Talat, "Directions for NLP practices applied to online hate speech detection" in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, Y. Zhang, Eds., (Association for Computational Linguistics, 2022), pp. 11794–11805.
12. P. Röttger *et al*., "Hatecheck: Functional tests for hate speech detection models" in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, C. Zong, F. Xia, W. Li, R. Navigli, Eds., (Association for Computational Linguistics, 2021), pp. 41–58.
13. D. W. Sue, Race talk: The psychology of racial dialogues. *Am. Psychol.* **68**, 663–672 (2013).
14. E. D. Knowles, B. S. Lowery, R. M. Chow, M. M. Unzueta, Deny, distance, or dismantle? How white Americans manage a privileged identity *Perspect. Psychol. Sci.* **9**, 594–609 (2014).
15. P. A. Goff, C. M. Steele, P. G. Davies, The space between us: Stereotype threat and distance in interracial contexts. *J. Pers. Soc. Psychol.* **94**, 91–107 (2008).
16. C. M. Steele, S. J. Spencer, J. Aronson, Contending with group image: The psychology of stereotype and social identity threat. *Adv. Exp. Soc. Psychol.* **34**, 379–440 (2002).
17. J. Guynn, Facebook while black: Users call it getting "Zucked," say talking about racism is censored as hate speech. *USA Today*, 9 July 2020. https://www.usatoday.com/story/news/2019/04/24/facebook-while-black-zucked-users-say-they-get-blocked-racism-discussion/2859593002/. Accessed 4 December 2023.
18. C. Silverman, Black lives matter activists say they're being silenced by facebook. Buzzfeed (2020). https://www.buzzfeednews.com/article/craigsilverman/facebook-silencing-black-lives-matter-activists. Accessed 4 December 2023.
19. A. Holmes, Black instagram users were 50 their accounts automatically disabled, internal research reportedly showed. *Business Insider* (2020). https://www.businessinsider.com/black-instagram-users-faced-disproportionate-bans-report-2020-7. Accessed 4 December 2023.
20. M. McCluskey, These tiktok creators say they're still being suppressed for posting black lives matter content. Time (2020). https://time.com/5863350/tiktok-black-creators/. Accessed 4 December 2023.
21. T. Mitchell, Black creators say tiktok's algorithm fosters a'consistent undertone of anti-blackness.'here's how the app has responded. Business Insider (2021). https://www.businessinsider.com/a-timeline-of-allegations-that-tiktok-censored-black-creators-2021-7. Accessed 4 December 2023.
22. O. L. Haimson, D. Delmonaco, P. Nie, A. Wegner, Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proc. ACM Hum. Comput. Interact.* **5**, 1–35 (2021).
23. P. C. Cozby, Self-disclosure: A literature review. *Psychol. Bull.* **79**, 73–91 (1973).
24. R. Lin, S. Utz, Self-disclosure on SNS: Do disclosure intimacy and narrativity influence interpersonal closeness and social attraction? *Comput. Hum. Behav.* **70**, 426–436 (2017).
25. R. M. Kowalski, Complaints and complaining: Functions, antecedents, and consequences. *Psychol. Bull.* **119**, 179–196 (1996).
26. E. Kubin, C. Puryear, C. Schein, K. Gray, Personal experiences bridge moral and political divides better than facts. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2008389118 (2021).
27. C. R. Kaiser, C. T. Miller, Stop complaining! The social costs of making attributions to discrimination *Pers. Soc. Psychol. Bull.* **27**, 254–263 (2001).
28. C. R. Kaiser, C. T. Miller, Derogating the victim: The interpersonal consequences of blaming events on discrimination. *Group Process. Intergroup Relat.* **6**, 227–237 (2003).
29. E. R. Carter, M. C. Murphy, Consensus and consistency: Exposure to multiple discrimination claims shapes whites' intergroup attitudes. *J. Exp. Soc. Psychol.* **73**, 24–33 (2017).
30. E. Bonilla-Silva, *Racism without racists: Color-Blind Racism and the Persistence of Racial Inequality in the United States* (Rowman & Littlefield Publishers, 2006).
31. M. W. Kraus, I. N. Onyeador, N. M. Daumeyer, J. M. Rucker, J. A. Richeson, The misperception of racial economic inequality. *Perspect. Psychol. Sci.* **14**, 899–921 (2019).
32. D. S. Massey, J. Rothwell, T. Domina, The changing bases of segregation in the United States. *Ann. Am. Acad. Polit. Soc. Sci.* **626**, 74–90 (2009).
33. D. S. Massey, N. A. Denton, "American apartheid: Segregation and the making of the underclass" in *Social Stratification, Class, Race, and Gender in Sociological Perspective, Second Edition*, D. Grusky, Ed. (Routledge, 2019).
34. D. Kumar, Y. AbuHashem, Z. Durumeric, Watch your language: Large language models and content moderation. arXiv [Preprint] (2023). http://arxiv.org/abs/2309.14517 (Accessed 8 April 2024).
35. C. Ziems *et al*., Can large language models transform computational social science?. *Comput. Linguist.* **50**, 237–291 (2024).
36. C. Hutto, E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text" in *Proceedings of the International AAAI Conference on Web and Social Media* (The Association for the Advancement of Artificial Intelligence, 2014), vol. 8, pp. 216–225.
37. J. W. Pennebaker, R. J. Booth, M. E. Francis, LIWC2007: Linguistic inquiry and word count (2007). https://www.liwc.net. Accessed 3 April 2023.
38. X. Zhou, M. Sap, S. Swayamdipta, Y. Choi, N. Smith, "Challenges in automated debiasing for toxic language detection" in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics* (Association for Computational Linguistics, 2021).
39. E. Dwoskin, N. Tiku, C. Timberg, Facebook's race-blind practices around hate speech came at the expense of black users, new documents show. *The Washington Post*, 21 November 2021. https://www.washingtonpost.com/technology/2021/11/21/facebook-algorithm-biased-race/. Accessed 4 December 2023.
40. R. Gorwa, R. Binns, C. Katzenbach, Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data Soc.* **7**, 205395171989794 (2020).

41. T. Davidson, D. Bhattacharya, I. Weber, "Racial bias in hate speech and abusive language detection datasets" in *Third Workshop on Abusive Language Online*, S. T. Roberts, J. Tetreault, V. Prabhakaran, Z. Waseem, Eds., (Association for Computational Linguistics, Florence, Italy, 2019), pp. 25–35.
42. T. Hartvigsen *et al*., "Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection" in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, A. Villavicencio, Eds., (Association for Computational Linguistics, Dublin, Ireland, 2022), pp. 3309–3326.
43. T. Davidson, D. Warmsley, M. Macy, I. Weber, "Automated hate speech detection and the problem of offensive language" in *Proceedings of the international AAAI conference on web and social media.*, (The Association for the Advancement of Artificial Intelligence, 2017), vol. 11, pp. 512–515.
44. K. Ethayarajh, Y. Choi, S. Swayamdipta, "Understanding dataset difficulty with v-usable information" in *International Conference on Machine Learning*, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, S. Sabato, Eds., (Proceedings of Machine Learning Research, 2022), pp. 5988–6008.
45. L. Dixon, J. Li, J. Sorensen, N. Thain, L. Vasserman, "Measuring and mitigating unintended bias in text classification" in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, (Association for Computing Machinery, 2018), pp. 67–73.
46. K. Gligoric, M. Cheng, L. Zheng, E. Durmus, D. Jurafsky, "NLP Systems That Can't Tell Use from Mention Censor Counterspeech, but Teaching the Distinction Helps" in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, K. Duh, H. Gomez, S. Bethard, Eds., (Association for Computational Linguistics, Mexico City, Mexico, 2024), pp. 5942–5959.
47. M. Sap *et al*., "Annotators with attitudes: How annotator beliefs and identities bias toxic language detection" in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, M. Carpuat, M.-C. de Marneffe, I. V. Meza Ruiz, Eds., (Association for Computational Linguistics, 2022), pp. 5884–5906.
48. Facebook, Community standards (2023). https://transparency.fb.com/policies/community-standards/. Accessed 4 December 2023.
49. H. Tajfel, Experiments in intergroup discrimination. *Sci. Am.* **223**, 96–103 (1970).
50. H. Tajfel, J. C. Turner, *The Social Identity Theory of Intergroup Behavior in Political Psychology* (Psychology Press, 2004), pp. 276–293.
51. M. B. Brewer, The importance of being we: Human nature and intergroup relations. *Am. Psychol.* **62**, 728–738 (2007).
52. N. R. Branscombe, N. Ellemers, R. Spears, B. Doosje, The context and content of social identity threat. *Soc. Identity Context Commitment Content* **1**, 35–58 (1999).
53. J. Zaki, Empathy: A motivated account. *Psychol. Bull.* **140**, 1608–1647 (2014).
54. F. Bernardo, J. M. Palma-Oliveira, Identification with the neighborhood: Discrimination and neighborhood size. *Self Ident.* **15**, 579–598 (2016).
55. J. F. Dovidio, T. Saguy, N. Shnabel, Cooperation and conflict within groups: Bridging intragroup and intergroup processes. *J. Soc. Issues* **65**, 429–449 (2009).
56. N. Miller, M. B. Brewer, *Categorization Effects on Ingroup and Outgroup Perception* (Academic Press, 1986).
57. I. Oluo, *So You Want to Talk About Race* (Hachette, 2019).
58. J. Kwarteng, S. C. Perfumi, T. Farrell, M. Fernandez, "Misogynoir: public online response towards self-reported misogynoir" in *Proceedings of the 2021 IEEE/ACM international conference on advances in social networks analysis and mining*, M. Coscia, A. Cuzzocrea, K. Shu, Eds., (Association for Computing Machinery, 2021), pp. 228–235.
59. A. Lorde, The uses of anger. *Women's Stud. Q.* **25**, 278–285 (1997).
60. B. L. Monroe, M. P. Colaresi, K. M. Quinn, Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Polit. Anal.* **16**, 372–403 (2008).
61. L. Freeman, H. Stewart, Toward a harm-based account of microaggressions. *Perspect. Psychol. Sci.* **16**, 1008–1023 (2021).
62. R. T. Lee, A. D. Perez, C. M. Boykin, R. Mendoza-Denton, On the prevalence of racial discrimination in the United States. *PloS One* **14**, e0210698 (2019).
63. A. Shahani, Twitter adds warning label for offensive political tweets. *NPR* (2019). https://www.npr.org/2019/06/27/736668003/twitter-adds-warning-label-for-offensive-political-tweets#:~:text=The%20company%20will%20not%20delete,behavior%20apply%20to%20this%20Tweet. Accessed 4 December 2023.
64. S. L. Gaertner, J. Mann, A. Murrell, J. F. Dovidio, Reducing intergroup bias: The benefits of recategorization. *J. Personal. Soc. Psychol.* **57**, 239–249 (1989).
65. K. H. Greenaway, R. G. Wright, J. Willingham, K. J. Reynolds, S. A. Haslam, Shared identity is key to effective communication. *Personal. Soc. Psychol. Bull.* **41**, 171–182 (2015).
66. J. L. Kalla, D. E. Broockman, Which narrative strategies durably reduce prejudice? Evidence from field and survey experiments supporting the efficacy of perspective-getting *Am. J. Polit. Sci.* **67**, 185–204 (2023).
67. J. S. Coke, C. D. Batson, K. McDavis, Empathic mediation of helping: A two-stage model. *J. Personal. Soc. Psychol.* **36**, 752–766 (1978).
68. A. R. Todd, A. D. Galinsky, Perspective-taking as a strategy for improving intergroup relations: Evidence, mechanisms, and qualifications. *Soc. Personal. Psychol. Compass* **8**, 374–387 (2014).
69. E. Kubin, C. von Sikorski, The role of (social) media in political polarization: A systematic review. *Ann. Int. Commun. Assoc.* **45**, 188–206 (2021).
70. C. A. Bail *et al*., Exposure to opposing views on social media can increase political polarization. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 9216–9221 (2018).
71. R. B. Cialdini, R. P. Jacobson, Influences of social norms on climate change-related behaviors. *Curr. Opin. Behav. Sci.* **42**, 1–8 (2021).
72. J. Kim, C. McDonald, P. Meosky, M. Katsaros, T. Tyler, Promoting online civility through platform architecture. *J. Online Trust Safety* **1** (2022).
73. J. A. Okonofua, L. T. Harris, G. M. Walton, Sidelining bias: A situationist approach to reduce the consequences of bias in real-world contexts. *Curr. Dir. Psychol. Sci.* **31**, 395–404 (2022).
74. Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
75. S. Santy, J. T. Liang, R. L. Bras, K. Reinecke, M. Sap, "NLPositionality: Characterizing Design Biases of Datasets and Models" in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, N. Okazaki, Eds., (Association for Computational Linguistics, Toronto, Canada, 2023), pp. 9080–9102.
76. B. Van Aken, J. Risch, R. Krestel, A. Löser, "Challenges for toxic comment classification: An in-depth error analysis" in *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, D. Fišer, R. Huang, V. Prabhakaran, R. Voigt, Z. Waseem, J. Wernimont, Eds., (Association for Computational Linguistics, Brussels, Belgium, 2018), pp. 33–42.
77. T. Garg, S. Masud, T. Suresh, T. Chakraborty, Handling bias in toxic speech detection: A survey. *ACM Comput. Surv.* **55**, 1–32 (2023).
78. C. Jia, M. S. Lam, M. C. Mai, J. Hancock, M. S. Bernstein, "Embedding democratic values into social media AIs via societal objective functions." in *Proceedings of the ACM on Human-Computer Interaction*, J. Nichols, Ed. (Association for Computing Machinery, 2024), pp. 1–36.
79. Q. Wu, B. Semaan, "how do you quantify how racist something is?": Color-blind moderation in decentralized governance *Proc. ACM Hum. Comput. Interact.* **7**, 1–27 (2023).
80. G. Weld, L. Leibmann, A. X. Zhang, T. Althoff, Perceptions of moderators as a large-scale measure of online community governance. arXiv [Preprint] (2024). http://arxiv.org/abs/2401.16610 (Accessed 10 April 2024).
81. L. P. Argyle *et al*., Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2311627120 (2023).
82. J. N. Matias, Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 9785–9789 (2019).
83. A. Kozyreva *et al*., Resolving content moderation dilemmas between free speech and harmful misinformation. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2210666120 (2023).
84. J. Grimmer, B. M. Stewart, Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Polit. Anal.* **21**, 267–297 (2013).
85. A. Field *et al*., "Framing and agenda-setting in Russian news: A computational analysis of intricate political strategies" in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii, Eds., (Association for Computational Linguistics, Brussels, Belgium, 2018), pp. 3570–3580.
86. J. Zhang, W. Hamilton, C. Danescu-Niculescu-Mizil, D. Jurafsky, J. Leskovec, "Community identity and user engagement in a multi-community landscape" in *Proceedings of the International AAAI Conference on Web and Social Media*, (The Association for the Advancement of Artificial Intelligence, 2017), vol. 11, pp. 377–386.
87. C. R. Harris *et al*., Array programming with Numpy. *Nature* **585**, 357–362 (2020).
88. S. Seabold, J. Perktold, "Statsmodels: Econometric and statistical modeling with python" in *Proceedings of the 9th Python in Science Conference* (SciPy, 2010). vol. 57.
89. C. Lee, Data from "Online suppression of racial discrimination disclosures." OSF. https://osf.io/f3eqt/. Deposited 20 March 2024.