Who Gets Flagged? An Experiment on Censorship and Bias in Social Media Reporting

Jessica T. Feezell, *University of New Mexico, USA*
Meredith Conroy, *California State University, San Bernardino, USA*
Barbara Gomez-Aguinaga, *Stanford University, USA*
John K. Wagner, *University of New Mexico, USA*

With a large majority of Americans using social media platforms to consume and disseminate information on a regular basis, social media serve as today's town square in many ways (Pew Research Center 2021). However, unlike public spaces where the free expression of citizens is afforded First Amendment protections, social media platforms are privately owned, and users are subject to the platform's terms of service and community standards (Congressional Research Service 2021). Although platform rules vary about what is allowable content, most are in agreement that certain forms of content (e.g., credible threats of violence and hate speech) are not, and they strive to identify and remove such posts. Both Twitter and Facebook prohibit credible threats of violence (e.g., "I will…" or "I plan to…") and hate speech directed at protected classes (e.g., race, gender, and religion). To identify objectionable content, social media platforms rely in part on users to report offensive posts, which the platform then decides to leave up or take down (Crawford and Gillespie 2016). Users play a critical role in determining which content is flagged for review; however, little is known about user reporting behavior.

In general, social media platforms use two techniques to identify objectionable content: (1) algorithms (or "classifiers") that are trained to flag posts that contain certain language; and (2) other users who report posts that they believe violate the community standards (Crawford and Gillespie 2016). Posts that are identified as possibly containing objectionable content then are reviewed by a group of human moderators to determine whether the post in fact violates the terms of service and therefore should be removed or labeled. Adjudicating what is and is not objectionable content is difficult and subject to personal biases; even professional moderators admit to making mistakes (Gadde and Derella 2020; Varner et al. 2017). However, classifiers also are subject to racial bias. For instance, several classifiers were more likely to flag social media posts written in "Black English" as abusive than posts written in standard English (Davidson, Bhattacharya, and Weber 2019; Sap et al. 2019). Automated toxic-language identification tools generally are unable to consider social and cultural context and therefore risk reporting posts that are not actually in violation. Thus, the assumption that automated techniques are a way to remove bias is incorrect and may invite systemic bias.

In our study, we tested for bias in the second pathway to online content removal: that is, through social media users. Specifically, we were interested in whether the demographics of the poster influence a willingness to report content as violating the community standards; this makes certain demographics more likely to have their posts reviewed and possibly removed. We focused on race, gender, and the intersection of these traits because gendered and racial stereotypes—as well as shared traits between messengers and receivers—can influence people's attitudes and evaluations of content (Karpowitz, Mendelberg, and Shaker 2012; Mastro 2017). Although some scholars argue that computer-mediated communication has reduced the public's ability to identify the background of messengers, other studies have shown that personal characteristics of the public continue to influence assessments of messages in online environments (Metzger and Flanagin 2013; Settle 2018).

DEMOGRAPHIC TRAITS AND ASSESSMENTS IN ONLINE ENVIRONMENTS

The Social Role Theory (SRT) contends that social norms and stereotypes can have significant implications for evaluations of minority groups and women, even when conducting the same activities or delivering the same messages (Diekman and Eagly 2000; Schneider and Bos 2019). Minority groups often are associated with specific tasks, activities, and behaviors that are perceived as expected or appropriate for a given group (Luisi, Adams, and Kilgore 2021; Negrón-Muntaner et al. 2014). In mass media, the absence of women in male-dominated fields or their infrequent appearance as experts has led to lower evaluations of women, even when they deliver the same message as men (Armstrong and McAdams 2009; Luisi, Adams, and Kilgore 2021). Moreover, according to SRT, men are rewarded for exhibiting behaviors that are congruent with an agentic personality and women are rewarded for exhibiting behaviors that are associated with communion—and likewise punished for those that are incongruent with

doi:10.1017/S1049096522001238

these expected personality profiles (Koenig and Eagly 2014). Similarly, the overrepresentation of African Americans as "lawbreakers" or "servants" and US Latinos as "unauthorized immigrants" or "non-English speakers" in the marketing, news, and entertainment industries is associated with negative assessments of these groups (e.g., lower levels of perceived credibility, leadership, and intelligence). Hence, the evaluation of the behavior or messages from minority groups tends to be contingent on existing norms and stereotypes.

Although users play a critical role in determining which content is flagged for review, little is known about user reporting behavior.

Building on SRT, the Social Information Processing Theory (SIPT) contends that the implications of social norms and the backgrounds of messengers and receivers also transfer to online settings (Marmat 2022). This occurs because the web environment provides users with cues to "perceive and evaluate their communication partners with a similar level of accuracy to that of face-to-face communicators by utilizing and accumulating whatever information is available within the environment" (Lee and Lim 2014, 556). Online cues include names, profile photographs, and videos, which allow computer-mediated communication participants to examine multiple sociodemographic characteristics of the messengers (both elites and the public), including sex, age, and racial and ethnic backgrounds (Settle 2018). Some of these traits have been shown to influence receivers' attitudes or evaluations not only in face-to-face interactions but also in printed and online environments. Based on SRT and SIPT, we expected that social media users would be more likely to report posts from women and ethnic minorities due to stereotypes associated with these groups (Hypothesis 1).

THE MEDIATING EFFECTS OF SHARED BACKGROUNDS

Shared demographic characteristics between messengers and receivers, however, can mediate the negative implications of strict social norms including gendered and racial stereotypes (Armstrong and McAdams 2009). In diverse contexts such as the United States, message receivers—particularly those from underrepresented groups—often acknowledge similarities between themselves and the message sources. These perceived similarities have the potential to influence thoughts, decisions, and evaluations because "we tend to compare ourselves to others, and the greater the similarity we perceive, the more likely we are to attend to and trust what [messengers] say" (Andsager and Mastin 2003, 59). In other words, shared demographic characteristics such as race and ethnicity can be salient in a social media user's mind when evaluating inflammatory content. As a result, we expected that users would be less likely to report posts from people with whom they share demographic characteristics (Hypothesis 2).

EXPERIMENTAL DESIGN

To examine whether some users are more likely to be reported than others for their inflammatory but rule-abiding content, we conducted an experiment embedded in the Knight Foundation–Ipsos Freedom of Expression Study, which was fielded online using the probability-based Ipsos KnowledgePanel on July 30, 2021, to a sample of 2,500 panel members (Feezell et al. 2022). Each subject was shown a short list of produced social media posts designed to mimic a mini-Twitter feed and they were asked to report any posts that they believed violated the terms of agreement (see online appendix A for examples). Specifically, the experiment pre-text read as follows:

> Social media platforms like Facebook and Twitter are governed by "terms of service" and "community standards" regarding the content they allow on their platform. Which of the following posts, if any, would you "report" as violations of community standards?

Participants then were exposed to four mini-feeds on separate pages, each featuring one inflammatory post and two decoy posts.[1] Each inflammatory post was designed to be provocative but not to violate the rules of common social media platforms, which prohibit threats of credible violence or hate speech directed at protected classes.[2] As a result, this experiment allowed us to examine the likelihood that a post was reported, which—in the real world—would subject the post to review up the chain of command and cause it to experience a higher probability of mistakenly being censored.

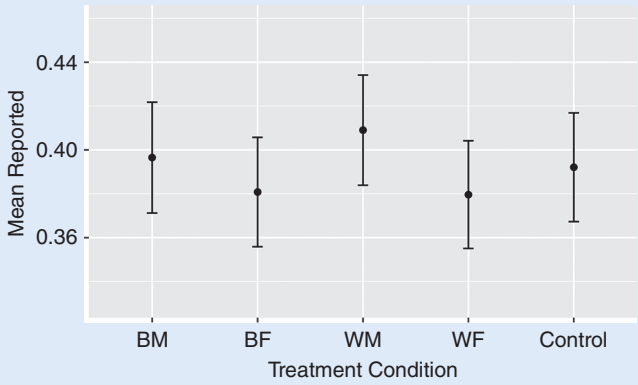
For our treatment, each inflammatory post was randomized to vary the gender (Male; Female) and race (White Non-Hispanic; Black or African American) of the poster through the photograph and name of the person making the comment.[3] This created a 2X2 factorial design plus a control group (i.e., Black Male, Black Female, White Male, White Female, Control). Subjects were asked to check a box next to any posts in the list that they would "report" for violating community standards. Online appendix B describes the factorial design.

To test Hypothesis 1, we compared the average reporting rate for each of the treatment conditions to the control condition across the four inflammatory posts. Averaging the four posts together allowed us to better isolate the influence of the treatment (i.e., race and gender) on the likelihood of reporting regardless of the idiosyncrasies of the content for each post. Figure 1 shows the mean values and 95% confidence intervals for each treatment condition compared to the control group.[4] Using four t-tests, we found that none of the treatment conditions were significantly different from the control group in their likelihood of posts being reported. To ensure that the findings were not biased by the wording of each post or the oversampling of young and diverse populations, we also ran four separate logistic regressions to predict whether respondents indicated that they would flag each post (0/1). We included a categorical variable for each experimental treatment as the main independent

https://doi.org/10.1017/S1049096522001238 Published online by Cambridge University Press

Figure 1

Effect of Treatment Assignment on Post Reporting



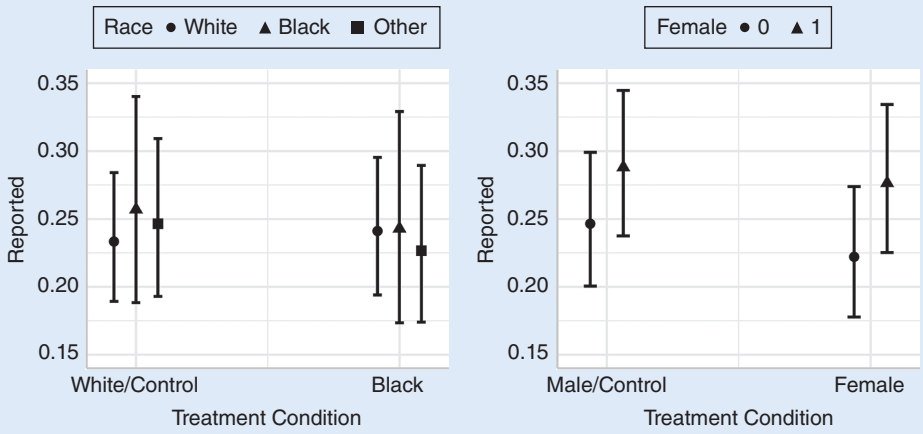
variable, which allowed for a comparison of each treatment group to the control group, as well as control variables for age, gender, race, income, education, and party identification (see online appendix C for descriptive statistics). Again, the treatment conditions were not significant predictors of post reporting when compared to the control group. We found through post hoc analysis, however, that when compared to Republicans, Democrats had a higher likelihood of reporting all posts ($p<0.05$ for all four posts) (see online appendix D for full models and appendix E for an analysis by partisanship).

To test Hypothesis 2, we examined the moderating role that the respondent's demographic characteristics play in deciding whether to report a post. We anticipated that people would be less likely to report others who they consider to be in their in-group. We ran a random-effects model interacting the respondent race and gender with the treatment condition race and gender to predict the probability that the respondent would report a post. Although it is likely that there is important nuance to be learned from more discrete categories of race and ethnicity, we restricted our analysis to concurrence between Black/White/Other respondents and Black/White treatment categories.

We ran separately two models with interactions for race and gender, including control variables for age, income, education, and party identification. Figure 2 (left side) presents the predicted probability of reporting a post comparing the race of the poster and the race of the participant. When interacting a post made by a Black poster with the participant's own race, there was no significant effect and thus no support for Hypothesis 2 concerning race. Individuals were no more or less likely to report a poster of the same race. Figure 2 (right side) presents the predicted probability of reporting a post comparing the gender of the poster and the gender of the participant. The interaction between poster gender and participant gender

Figure 2

Predicted Probability of Reporting a Post



Note: Bars show 95% confidence intervals.

https://doi.org/10.1017/S1049096522001238 Published online by Cambridge University Press

was not significant. Thus, whereas women were more likely to report posts across both models, they did not appear to be more likely to report a post based on the perceived gender of the poster (see online appendix F for full models). Therefore, we found no support for Hypothesis 2. Individuals were no more or less likely to report a poster based on shared demographic characteristics.

CONCLUSION

Our study examined whether the demographic characteristics of social media posters might render them more likely to be reported for violating the platform's terms of agreement. We expected that people of color and women would be more likely to have their posts reported for violation, but we did not find that in the study results. We also expected that being co-gender or co-racial with the poster would lead to less reporting, but we did not find those results either. Overall, we must accept the null for both of the hypotheses that we proposed.

In post hoc analysis that we did not theorize about in this study, we found that differences emerge when examining the backgrounds of social media users. Democrats, for example, were consistently more likely to report posts when compared to Republicans across all four posts (see online appendix D). Additionally, women were more likely to report than men in two of the four posts (in posts about law enforcement and Antifa). Older people were more likely to report posts in two of the four posts as well (in posts about a House Representative and law enforcement). Although our hypotheses about how the race and gender of the poster might lead to higher rates of reporting did not manifest, it was clear that certain segments of the user population are more likely than others to flag posts as inappropriate—regardless of the racial or gender background of the source.

This study raises questions about social media users and their likelihood to report the posts of women and ethnic minorities due to persisting stereotypes associated with these groups (Luisi, Adams, and Kilgore 2021; Negrón-Muntaner et al. 2014); however, there are limitations that future studies should address. Experimental approaches allow researchers to maximize internal validity by manipulating racial and gender cues. Although these cues are manageable in text-based experiments, the nature of social media settings presents additional challenges. Studies on first impressions from images have shown that facial expressions and image-capture conditions can influence the assessments of social media users (Todorov and Porter 2014). Given that we used stock images to complement the racial and gender cues from names, our study inadvertently may have captured additional biases influenced by the selection of the images. Another limitation of the study relies on disparities between the sample and the universe of social media users. The median age of the survey participants, for example, was 49.5, whereas the median age of Twitter users is 40 (Wojcik and Hughes 2019). This disparity reflects the possibility that the experiment did not rely on a representative sample of social media users, raising concerns about the external validity of the results. Finally, this study relied on prompts of groups with different social constructions, such as police officers, terrorists, and elected officials. Given that social constructions of professions and groups influence public opinion (Schneider and Ingram 1993), future research should design studies to compare the influence of distinct social constructions on reporting inflammatory posts.

SUPPLEMENTARY MATERIALS

To view supplementary material for this article, please visit http://doi.org/10.1017/S1049096522001238.

CONFLICTS OF INTEREST

The authors declare no ethical issues or conflicts of interest in this research. ■

NOTES

1. We used four different inflammatory posts to include ideologically diverse statements in the experiment and guard against introducing any bias that might result from interacting race and gender with ideology.
2. Inflammatory Post 1: "With all due respect, Representative, you are a liar, a rapist, and you should burn in hell." Inflammatory Post 2: "Terrorists deserve every form of torture—waterboarding is just the beginning." Inflammatory Post 3: "I don't care if you work security at Walmart, ALL cops are racist, the system is racist, and we should burn it all down." Inflammatory Post 4: "All these entitled Antifa pricks have never worked an honest day in their life. We'd be better off if they were all dead."
3. The treatment photographs were collected from a royalty-free photo website by searching for "Black male/female" and "white male/female." As a manipulation check, we conducted a convenience sample survey using the photographs and asked respondents to categorize them according to race and gender (N = 37). The race of the poster was categorized correctly 94% of the time and the gender of the poster was categorized correctly 93% of the time.
4. For each of the five conditions of the experiment—that is, the poster of the message being a Black Male, Black Female, White Male, White Female, or no identification—participants were asked about one to four posts from a given condition due to randomization of the treatment. Thus, the dependent variable we were analyzing was the proportion of posts reported by each participant for a given condition—not the overall proportion with which all participants reported a post of a specific condition. For example, Participant A may have seen four posts from a White Female and Participant B may have seen two. Participant A reported two posts and Participant B reported one. To avoid overrepresenting Participant A—as we would if we had calculated an overall proportion of White Female posts reported—we calculated that Participant A reported 50% of White Female posts and Participant B also reported 50%. Thus, to analyze differences between conditions, we had to compare the mean proportion of each condition with the mean proportion reported in our control condition. Because we then were comparing means, a t-test became the most appropriate statistical test.

REFERENCES

Andsager, Julie L., and Teresa Mastin. 2003. "Racial and Regional Differences in Readers' Evaluations of the Credibility of Political Columnists by Race and Sex." *Journalism & Mass Communication Quarterly* 80 (1): 57–72.

Armstrong, Cory L., and Melinda J. McAdams. 2009. "Blogs of Information: How Gender Cues and Individual Motivations Influence the Perceptions of Credibility." *Journal of Computer-Mediated Communication* 14 (3): 435–56.

Congressional Research Service. 2021. *Social Media: Misinformation and Content Moderation Issues for Congress.* Washington, DC: Congressional Research Service.

Crawford, Kate, and Tarleton Gillespie. 2016. "What Is a Flag For? Social Media Reporting Tools and the Vocabulary of Complaint." *New Media & Society* 18 (3): 410–28.

Davidson, Thomas, Debasmita Bhattacharya, and Ingmar Weber. 2019. "Racial Bias in Hate Speech and Abusive Language Detection Datasets." Florence, Italy: *Proceedings of the Third Workshop on Abusive Language Online*: 25–35.

Diekman, Amanda B., and Alice H. Eagly. 2000. "Stereotypes as Dynamic Constructs: Women and Men of the Past, Present, and Future." *Personality and Social Psychology Bulletin* 26 (10): 1171–88.

Feezell, Jessica T., Meredith Conroy, Barbara Gomez-Aguinaga, and John K. Wagner. 2022. "Replication Data for 'Who Gets Flagged? An Experiment on Censorship and Bias in Social Media Reporting.'" *PS: Political Science & Politics* DOI: 10.7910/DVN/WM6AOP.

Gadde, Vijaya, and Matt Derella. 2020. "An Update on Our Continuity Strategy During COVID-19." https://blog.twitter.com/en_us/topics/company/2020/An-update-on-our-continuity-strategy-during-COVID-19.

Karpowitz, Christopher F., Tali Mendelberg, and Lee Shaker. 2012. "Gender Inequality in Deliberative Participation." *American Political Science Review* 106 (3): 533–47.

Koenig, Annie M., and Alice L. Eagly. 2014. "Evidence for the Social Role Theory of Stereotype Content: Observations of Groups' Roles Shape Stereotypes." *Journal of Personality and Social Psychology* 107 (3): 371–92.

Lee, Jayeon, and Young-Shin Lim. 2014. "Who Says What About Whom: Young Voters' Impression Formation of Political Candidates on Social Networking Sites." *Mass Communication & Society* 17 (4): 553–72.

Luisi, Tim, Kelly L. Adams, and LaShawnda Kilgore. 2021. "Roughing the Caster! Sexism and Perceived Female Sports Broadcasters Credibility." *Atlantic Journal of Communication* 29 (4): 262–74.

Marmat, Geeta. 2022. "Online Brand Communication and Building Brand Trust: Social Information Processing Theory Perspective." *Global Knowledge, Memory, and Communication* 71 (6/7): 584–604.

Mastro, Dana. 2017. "Race and Ethnicity in US Media Content and Effects." *Oxford Research Encyclopedia of Communication*. https://doi.org/10.1093/acrefore/9780190228613.013.122.

Metzger, Miriam J., and Andrew J. Flanagin. 2013. "Credibility and Trust of Information in Online Environments: The Use of Cognitive Heuristics." *Journal of Pragmatics* 59 (B): 210–20.

Negrón-Muntaner, Frances, Chelsea Abbas, Luis Figueroa, and Samuel Robson. 2014. *The Latino Media Gap: A Report on the State of Latinos in U.S. Media.* www.scribd.com/document/230135450/Latino-Media-Gap-Report-by-Frances-Negron-Muntaner-with-Chelsea-Abbas-Luis-Figueroa-and-Samuel-Robson.

Pew Research Center. 2021. "*Social Media Fact Sheet.*" Washington, DC: Pew Research Center. www.pewresearch.org/internet/fact-sheet/social-media.

Sap, Maarten, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. "The Risk of Racial Bias in Hate Speech Detection." Florence, Italy: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*: 1668–78.

Schneider, Anne, and Helen Ingram. 1993. "Social Construction of Target Populations: Implications for Politics and Policy." *American Political Science Review* 87 (2): 334–47.

Schneider, Monica C., and Angela L. Bos. 2019. "The Application of Social Role Theory to the Study of Gender in Politics." *Political Psychology* 40 (S1): 173–213.

Settle, Jaime E. 2018. *Frenemies: How Social Media Polarizes America.* Cambridge: Cambridge University Press.

Todorov, Alexander, and Jenny M. Porter. 2014. "Misleading First Impressions: Different for Different Facial Images of the Same Person." *Psychological Science* 25 (7): 1404–17.

Varner, Madeleine, Ariana Tobin, Julia Angwin, and Jeff Larson. 2017. "*What Does Facebook Consider Hate Speech?*" New York: ProPublica. https://projects.propublica.org/graphics/facebook-hate.

Wojcik, Stefan, and Adam Hughes. 2019. "*Sizing up Twitter Users.*" Washington, DC: Pew Research Center. www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users.

https://doi.org/10.1017/S1049096522001238 Published online by Cambridge University Press