

Who Can Say What? Testing the Impact of Interpersonal Mechanisms and Gender on Fairness Evaluations of Content Moderation

Social Media + Society
October-December 2024: 1–15
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20563051241286702
journals.sagepub.com/home/sms
S Sage

Ina Weber^{1,2} , João Gonçalves², Gina M. Masullo³,
Marisa Torres da Silva⁴ , and Joep Hofhuis⁵ 

Abstract

Content moderation is commonly used by social media platforms to curb the spread of hateful content. Yet, little is known about how users perceive this practice and which factors may influence their perceptions. Publicly denouncing content moderation—for example, portraying it as a limitation to free speech or as a form of political targeting—may play an important role in this context. Evaluations of moderation may also depend on interpersonal mechanisms triggered by perceived user characteristics. In this study, we disentangle these different factors by examining how the gender, perceived similarity, and social influence of a user publicly complaining about a content-removal decision influence evaluations of moderation. In an experiment ($n = 1,586$) conducted in the United States, the Netherlands, and Portugal, participants witnessed the moderation of a hateful post, followed by a publicly posted complaint about moderation by the affected user. Evaluations of the fairness, legitimacy, and bias of the moderation decision were measured, as well as perceived similarity and social influence as mediators. The results indicate that arguments about freedom of speech significantly lower the perceived fairness of content moderation. Factors such as social influence of the moderated user impacted outcomes differently depending on the moderated user's gender. We discuss implications of these findings for content-moderation practices.

Keywords

content moderation, online hate, free speech, gender stereotypes, social influence

Introduction

Social media platforms, once seen as spaces for connecting people, have come under scrutiny for their role in the proliferation of harmful content and the reinforcement of stereotypes. For instance, women and other marginalized groups are more likely to face harsh and sometimes biased scrutiny as content creators on social media (Gerrard & Thornham, 2020; Haimson et al., 2021). They are also more likely to become targets of online harassment when posting about their work in journalism or politics (Park et al., 2023). Content policies serve the purpose of protecting vulnerable groups from discrimination while ensuring that freedom of speech is not disproportionately limited (Riedl et al., 2022). However, researchers have identified several asymmetries in their enforcement that favor the viewpoints of majority groups (Pettersson, 2019).

On mainstream social media, the line that separates acceptable from unacceptable speech is determined by few technology companies but influences public debate and the interactions of millions of users worldwide. User perspectives

on content moderation and whether it complies with expected standards such as fairness and impartiality (Suzor et al., 2018) are therefore relevant to be examined more prominently in research on content moderation (Riedl et al., 2022). Gender as a salient characteristic is known to play a crucial role in online interactions (Wilhelm, 2020) while also impacting the perceived appropriateness of behaviors and emotion displays (Prentice & Carranza, 2002). Therefore, this study investigates gender as a factor that could drive support or rejection of moderation decisions on social media.

¹University of Antwerp, Belgium

²Erasmus University Rotterdam, The Netherlands

³The University of Texas at Austin, USA

⁴Universidade Nova de Lisboa, Portugal

⁵Amsterdam University of Applied Sciences, The Netherlands

Corresponding Author:

Gina M. Masullo, The University of Texas at Austin, 300 W. Dean Keeton St., Austin, Texas, USA 78712.

Email: gina.masullo@austin.utexas.edu



How content-moderation decisions are perceived needs to be understood within the context of broader debates on the topic. The most prominent counterargument invokes the importance of free speech, as illustrated in the case of billionaire Elon Musk's purchase of the platform formerly known as Twitter. He used the protection of free speech as an argument for scaling back moderation (Robertson, 2022) and reinstating the accounts of former U.S. president Donald Trump and rapper Kanye West, banned for inciting violence and explicit anti-Semitism, respectively (Mac & Browning, 2022; Rothenberg, 2023).

Expressive acts may be tolerated more when they are performed by certain favored groups, while general rejections independent of the actor only occur in cases of highly deviant and broadly condemned forms of opinion expression (e.g., burning a country's flag; Hurwitz & Mondak, 2002). Considering the importance of group identification, as well as status and influence in online contexts, we assume that these factors are crucial for better understanding user evaluations of content moderation (Hogg, 2021). While studies considered how users make sense of the technologies and reasonings employed in the moderation process (Gonçalves et al., 2023; Myers West, 2018), it still remains unclear which mechanisms may underly and explain evaluations of content moderation.

This study explores social influence and perceived similarity as two interpersonal mechanisms that may hold significant influence in this context. For example, offensive posts made by users who present themselves as championing free speech may be more readily accepted by others who value freedom of expression because they see themselves as similar to the users posting the offensive content. We assess the impact of these mechanisms in an experiment testing whether the gender of another user affects perspectives on the fairness, legitimacy, and bias of a moderation decision. We also consider the impact of public complaints against a moderation decision, referencing free speech or political targeting of conservatives as arguments commonly raised to discredit content moderation. The study was conducted in the United States, the Netherlands, and Portugal, allowing to account for nuanced differences between cultural contexts and informing further comparative work on factors that affect perceptions of content moderation.

With the selection of three Western consolidated democracies with high levels of internet usage, we aimed to maintain similar contexts while maximizing variability in the perceptions of online hate and gender roles. There are known differences between the United States and the European Union in terms of policy (minimal vs. highly developed legal frameworks on hate speech) and sensitivities toward what is considered hateful (Weber et al., 2020). Meanwhile, variability also exists between European countries regarding the degrees to which hateful speech is found tolerable in Southern and Northern regions (Celuch et al., 2022). Despite all ranking highly in global comparisons of

gender equality, combined indices on women's health, empowerment, and labor market participation place the Netherlands in the 5th, Portugal in the 21st, and the United States in the 44th rank of gender equality (United Nations Development Programme [UNDP], 2024).

Results of this study revealed that arguments about free speech significantly lower the perceived fairness of moderation decisions among bystanders. User gender was connected to higher levels of perceived social influence for women, leading to more negative evaluations of content removals. The study, thus, depicts a complex set of factors that shape how social media users perceive content-moderation decisions. These insights hold implications for how content moderation as an organizational practice against harmful content can be depicted more favorably but also informs on-going critical debates about what can be said online and by whom.

Literature Review

Evaluating the Fairness of Content Moderation

The term content moderation can be used to refer to a range of both top-down and community-driven practices (Jhaver, Appling, et al., 2019). This study is concerned with content moderation as a practice executed by social media platforms with the intention of limiting the spread of harmful user-generated content (UGC). To this end, platforms commonly draft content policies to define which types of content are not permissible and to enforce their policies by relying on automated systems, human content moderators, or a combination of both to screen UGC and remove posts or comments if policies are violated (Shaughnessy et al., 2024). Consequently, moderation on social media is not only a response to problematic content and behaviors. It is a practice deeply embedded into the organizational and commercial structures of companies whose primary service is the provision of platforms for the production, dissemination, and consumption of UGC and has implications for users in different roles (e.g., content creators and consumers; Shaughnessy et al., 2024).

This approach toward content moderation makes organizational justice a suitable framework for conceptualizing how moderation is evaluated by users. Organizational justice understands fairness as a product of fair process and fair outcome, executed by a legitimate authority (Colquitt, 2001; Colquitt et al., 2013). Legitimacy implies that an entity such as a moderator is accepted as having the right to make decisions about content on a platform (van Dijke et al., 2010). This right derives from principles akin to those of the rule of law, such as consent from users, equality and predictability, and due process (Suzor et al., 2018). Our study touches upon these directly: on the perceived fairness of the process, gender equality in treatment, and how user complaints as a form of public appeal against a moderation decision relate to due process. Perceived bias additionally describes the extent to which moderation is seen as exhibiting preferential treatment

of one side. Positive evaluations of content moderation thus view it as a fair, legitimate, and unbiased practice.

Situational cues can make deviant expressions more acceptable in certain cases. For instance, offensive words can be reclaimed for empowerment when appropriated by the groups they aimed to offend (Stollznow, 2020), and online calls for the death of Russian soldiers were found acceptable in the context of the war in Ukraine (Trevelyan, 2022). When individuals can be expected to tolerate public expressions of offensive or hateful speech in some cases, there may be variation in how appropriate they find the removal of content posted by other users on social media. As studies show, bystander support for content moderation depends not only on the content itself but also on the platform and its guidelines, the moderator, and the transparency of the process (Jhaver, Bruckman, & Gilbert, 2019; Suzor et al., 2019). Detailed explanations from platforms about a moderation decision can enhance bystander support for content moderation (Gonçalves et al., 2023). However, explicit justifications are rarely provided, prompting bystanders to rely on other contextual information to evaluate moderation decisions.

On most major social media platforms, users affected by content removals can use a formalized system to appeal a moderation decision. While these nonpublic appeal mechanisms are designed to mitigate feelings of being treated unfavorably, there is ample indication that appeal procedures often create frustration among users, for example, due to a lack of reactivity from the platform (Myers West, 2018; Vaccaro et al., 2020). Subsequently, some users may choose to express their discontent about a moderation decision publicly. Because explanations for moderation decisions can enhance bystander support, publicly posted complaints could have the opposite effect and sway opinions in the other direction. We thus propose that:

Hypothesis 1 (H1). A public user complaint about a moderation decision influences participants to perceive moderation of a hateful post by another user more negatively (less fair, less legitimate, more biased) than when no complaint is made.

Interpersonal Mechanisms: Perceived Similarity and Social Influence

On social media, cues about an individual's identity or group belonging form crucial building blocks for social relations. Followers of social media influencers for example feel more readily connected to—and receptive for product promotions by—influencers who they perceive as sharing identity-based similarities with themselves (Gupta et al., 2023). In the context of political convictions, study participants in the United States were found to assess the value of a policy proposal based on the party that allegedly proposed it, regardless of whether its content aligned with or opposed their own political stance (Van Bavel & Packer, 2021). With identities being

increasingly important to determine how politics are framed, they may also determine what kind of content is seen as acceptable to be publicly posted online. If what is admissible on social media is determined more by who says it, rather than the content itself, this could threaten principles of equality and openness in democracy, with far-reaching implications ranging from increased exclusion of marginalized groups to hateful content being widely disseminated just because it comes from influential individuals. We propose identities based on perceptions of similarity as a key mechanism that accounts for the impact of a public complaint and assume that the degree of perceived similarity between a user complaining about a moderation decision and a bystander might influence reactions to content moderation. This assumption is rooted in psychological processes related to social identification, which play a role in the dynamics of online interactions (Croes & Bartels, 2021).

Social identity theory explains intergroup behaviors based on perceptions of group membership, group differences, and issues of legitimacy regarding group status differences (Tajfel et al., 1979). While ingroups and outgroups may be clearly delineated in many real-life settings, online environments often afford a more fluid interpretation of intergroup boundaries. In fleeting interactions on social media, it may be unfeasible for users to assign each other's group membership due to a lack of factual information. Instead, users may use the limited available information and assign group membership based on stereotypes or prototypicality of the characteristics in question. If perceived similarity with another user is high, the other is subconsciously categorized as ingroup member, priming ingroup favoritism. If perceived as dissimilar, other users are categorized as outgroup members, priming outgroup derogation (Rathje et al., 2021).

Perceived similarity can make individuals more tolerant toward deviant behavior (Hurwitz & Mondak, 2002). A so-called “deviance credit” for instance legitimizes certain behaviors when performed by ingroup members (Abrams et al., 2018). Thus, in fleeting and non-recurring online interactions, limited available cues may lead bystanders to identify with other users subconsciously, by assessing perceived similarity. Applying the same reasoning to this study, we may expect that perceiving another user as similar to oneself makes individuals less prone to support deletion of that other user's content, even when the moderated post clearly represents a norm violation.

While social and nonverbal cues indicative of ingroup/outgroup belonging (e.g., age, race, class) are salient in offline contexts, such characteristics may be expressed online through public posts, shared content, or the connections of a user (Hasler & Amichai-Hamburger, 2013; Walther et al., 2011). Hence, an argument made in a social media post can indicate group belonging through shared ideas and opinions. In this study, two common arguments against content-moderation practices are used to indicate group belonging. Opponents of moderation often frame it as illegitimate censorship that

harms with free speech (Pettersson, 2019); thus, one argument endorses free speech while portraying moderation as a wrongful limitation to a fundamental right. Social media is often criticized for lacking political neutrality in moderation practices (Vogels et al., 2020). Conservative or right-leaning individuals often claim to be targeted on the platforms for their political viewpoints (Vaidhyanathan, 2019), which we address in the political persecution argument.

When raising these arguments in a public user complaint about a moderation decision, abstract collective identities (Hogg, 2005) such as being a free speech proponent or a victim of politically motivated mistreatment can be evoked. Both arguments can be considered partisan and may trigger a degree of perceived similarity between a bystander and another user, based on the perceived prototypical group membership of the user raising the argument (West & Iyengar, 2022). Although both arguments can appear in conjunction, they may still resonate differently as distinct collective ideas. We assume that, due to their relevance in the context of content moderation (Pettersson, 2019; Riedl et al., 2022), the salience of these arguments leads to attitudinal change about moderation (Althaus & Coe, 2011).

In their work on social identification processes, Postmes et al. (2005) show how inferred similarity on political opinions plays a role in small group dynamics. However, they also connect these social identities to mechanisms of social influence, under which group identities have a normative impact (e.g., support for free speech is incompatible with support for content moderation) and make other individuals appear as influential “prototypical representatives of the group” (Postmes et al., 2005, p. 6). This conceptual connection suggests social influence as another important mechanism that could account for the impact of public user complaints on evaluations of moderation.

Social influence describes an individual’s ability to exert influence on others in interactions and bring forth change in their attitudes, emotional states, or behaviors (Rashotte, 2007; Raven, 2008). For instance, a public statement on environmental action by a celebrity such as Taylor Swift would more likely produce attitude changes toward the environment than any form of UGC posted by a random individual on social media. This perspective understands social influence as a perceived characteristic of a person (Gnambs & Batinic, 2013). Higher social status and expectations of competence are related to higher perceived influence in social settings (Melamed & Savage, 2016; Oldmeadow et al., 2003). Applied to the context of this study, public user complaints against a moderation decision may increase identification through perceived similarity, causing people to perceive the outspoken user who denounced moderation as more socially influential. We expect this process to account for negative evaluations of a moderation decision. Assuming that the interpersonal mechanisms of perceived similarity and social influence play a mediating role, we propose the following hypothesis:

Hypothesis 2 (H2). The effect of H1 is mediated by participants’ perceived similarity with another user whose post was moderated (H2a) and by participants’ perceptions of another user whose post was moderated as having social influence (H2b).

Gender as a Cue for Evaluating Content Moderation

Gender deeply affects user experiences in digital environments and—particularly as a perceived characteristic on social media—shapes the interactions of users and how they conduct their own behavior (Wilhelm, 2020). This is most apparent within the issue of misogynistic online hate and harassment, targeting women in both personal and professional spheres and causing psychological harm as well as self-censoring tendencies (Park et al., 2023). The reproduction and reinforcement of gender norms moreover affects women as content creators, as evidenced by differential treatment upon posting controversial content (Gerrard & Thornham, 2020) or obstacles to access online communities for which women are stereotypically perceived as not skilled or competent enough (e.g., gamer communities or Wikipedia editors [Wagner et al., 2015]).

Most mainstream platforms are semi-anonymous, allowing for another user’s gender to be inferred from username, avatar, or pronouns and used as a cue for appraising them (Guegan et al., 2016). Appraisals of others connected to gender are tied to culturally shaped stereotypical notions about masculine or feminine traits and behaviors (e.g., confidence and ambition are associated with men, and warmth and care with women; Biernat & Sesko, 2018). These gender-specific stereotypes give indications on the social desirability and appropriateness of behavior and are divided into prescriptive characteristics (generally desirable but more for women or men) and proscriptive characteristics (generally undesirable but less for women or men; Prentice & Carranza, 2002). Behaviors lacking desirability and typicality are viewed particularly negative for violating prescriptive and conforming to proscriptive stereotypes (Wilhelm & Joeckel, 2019).

Dominant behavior, that is, exhibiting aggressiveness and competitiveness, is seen as typically masculine and generally more tolerated or even expected to be performed by men. Conversely, the same behaviors are sanctioned harshly when performed by women, who are expected to use a warm, communal style of communication (Carli, 2001). A woman’s dominant or aggressive behavior, thus, is evaluated in light of socio-cognitive expectations of gendered role behavior in different contexts (Prentice & Carranza, 2002). In the workplace for example, women who express anger are awarded less competence and lower status than men, for whom expressions of aggressiveness are tolerated as typical forms of emotional display (Brescoll & Uhlmann, 2008).

Our study focuses on the perceived gender of a user posting content. In the context of moderation, researchers found

that hateful comments were more likely to get flagged when the user posting them was a woman. Under consideration of emotion stereotypes, posting hateful content could be seen as engaging in both a general and a gender-specific norm violation for women, turning the breach of community guidelines into an act of “double deviance” (Wilhelm & Joeckel, 2019, p. 4). In online contexts with little available information about others, such stereotypical perceptions can be intensified (Guegan et al., 2016), potentially fostering disagreement with a hateful post by a woman and evoking positive attitudes toward its removal:

Hypothesis 3 (H3). Moderation of a hateful post by a woman will be perceived more positively by participants than moderation of a man’s hateful post.

Gender and emotion stereotypes were also connected to social influence in previous research (Salerno et al., 2019; Salerno & Peter-Hagene, 2015). As we assume perceived social influence to contribute to bystander evaluations of content-removal decisions, we explore the connection of this mechanism in relation to gender as a perceived characteristic. Men are usually assigned more social influence than women, especially when using dominant, negative communication styles (Salerno et al., 2019; Salerno & Peter-Hagene, 2015). According to language expectancy theory, men can use a wider range of emotions without losing persuasiveness (Burgoon et al., 2002). Differing preconceptions of appropriate communication related to gender and social status affect how much credibility a person is awarded (Craciun & Moore, 2019). Hence, aggressiveness is not only more tolerated when performed by men (Carli, 2001) but can even increase perceived competence and influence (Brescoll & Uhlmann, 2008). In group decision-making, contributions expressed with anger by men had a greater impact on outcomes than the same argument and emotion expressed by women. This was attributed to increased social influence of angry men, compared to a decrease in influence for angry women (Salerno et al., 2019; Salerno & Peter-Hagene, 2015).

Differences in perceived social influence vary across discussion settings and topics and are influenced by the salience of gender as an indicator of social status (Carli, 2001). Salerno and colleagues (2019) and Salerno and Peter-Hagene (2015) indicated gender only through names in a text-based interaction, which resembles the limited cues about gender on social media. Hence, we assume that women making a hateful post are perceived as aggressive and socially less influential:

Hypothesis 4 (H4). The effect of H2 is mediated by perceived social influence. A woman making a hateful post will be assigned less social influence, predicting more positive evaluations of a moderation decision regarding her post.

Evaluations of fairness, legitimacy, and bias of content moderation are likely shaped by cultural contexts and difficult to generalize. This may be connected to differences in

the tolerance toward online hate (Celuch et al., 2022), as well as the diverging centrality of freedom of speech in U.S. and European legal traditions (Heller & van Hoboken, 2019). These differences affect how hate speech is defined and persecuted, but also how public opinion toward the concept is formed (Bleich, 2011). However, EU countries are not homogeneous in this regard: Polls showed particularly high appreciation of free speech as a fundamental value in the Netherlands, while the absence of online speech regulations received low approval in Portugal (Naab, 2012).

There is also variation in how socio-cultural contexts shape perceptions of gender-related stereotypes. Although these are nuanced differences, Portuguese people appear more attached to traditional gender roles in household and family (47%) than the Dutch (15%; European Commission, 2017). Similarly, women in the United States are seen as facing more pressure regarding their role in parenting than men, for whom success at work and financial support of the family are seen as priorities (Pew Research Center, 2017). While not providing conclusive evidence for assuming country-specific differences, these insights emphasize that generalizability and robustness of our study’s findings benefit from cross-country perspectives.

Method

Data for this study were collected through Dynata in June 2020.¹ The final sample ($n=1,586$) was representative of internet users in the United States, the Netherlands, and Portugal for age, gender, and education, as well as race/ethnicity in the United States.² Detailed compositions of each sample are reported in the Supplemental Material.³ The study was approved by institutional ethics review boards in Europe and the United States and pre-registered (https://osf.io/qngfp/?view_only=2270c40915be408a9b6d5f87f56d2410). Data collection and analyses were conducted exactly as pre-registered, but for conciseness, our reporting focuses on main and mediation effects. Further analyses and point-by-point explanations for their exclusion from the main article are given in the Supplemental Material.

Stimuli, Study Design, and Procedures

The experiment used a 2 (user gender: man vs. woman) $\times$ 3 (user complaint: no complaint vs. free speech vs. political persecution) between-subjects design. Thereby, *user gender* refers to the gender of a fictitious social media user (as opposed to *participant gender*, which refers to demographic data). User complaint refers to a publicly posted statement contesting a content-moderation decision. Participants were first exposed to a hateful post by the fictitious user which contained a discriminatory statement against a minority group. Afterwards, participants were informed about the post’s removal by a content moderator for violating community standards. In the experimental conditions, participants were then shown another post in which the fictitious user complained about the content

removal. Depending on the condition, the complaint framed content moderation either as a limitation to free speech or as a form of political persecution.

All posts and notifications were fabricated and aimed to re-create a scenario that is likely to occur on any social media platform moderating content after it has been posted. On such platforms, users may encounter situations in which posts they previously saw on their social feed or another user's profile are removed due to violations of guidelines. They may also come across reactions such as a public complaint about this removal by the affected other user. The stimuli prominently displayed the name and profile photo of the fictitious user. Other visual cues were kept to a minimum to avoid distractions from the manipulations. The moderation message gave a brief notification about the removal of the post by a moderator without further details on the reasons why. It was kept consistent across all conditions.

With the aim of recreating a generic digital space for posting and discussing content, our stimuli were based on the style of comment sections of online news outlets using Disqus, a service for hosting and implementing commenting functions for websites. Figures A1–A4 depict examples of the stimuli. An overview of all stimuli messages and translations can be found in the Supplemental Material. In the survey, the posts were described as appearing on an unspecified social media platform. Through these design choices, we intended to create a realistic yet neutral environment. Although using the design of a popular platform such as Facebook or Twitter/X would increase external validity, it may also elicit preconceived notions about the types of users, content, and moderation practices. This could affect results and deter from the aim of detecting baseline effects across platform contexts.

All materials were composed in English and translated to Dutch and Portuguese contexts by native speakers. We ensured external validity by choosing typical names for the fictitious user and referring to a salient minority group in the hateful post (Mexican immigrants in the United States, refugees in the Netherlands, Brazilian immigrants in Portugal). The content of the hateful post was determined in a pre-test for a related study in all three countries ($n=304$), which was conducted to examine differences in the perception of various forms of offensive online content. The selected statement was perceived as distinctly more hateful than others (Gonçalves et al., 2023). The profile picture of the fictitious user was AI-generated and obtained from Generated Photos.⁴ To ensure that high or low levels of attractiveness of the profile picture do not affect how the fictitious user is perceived, another pre-test ($n=83$; 59% women) compared pictures of four different faces per user gender (Wang et al., 2010). The selected pictures were chosen according to lowest variance of attractiveness (man $M=4.65$, $SD=1.15$, $s^2=1.33$; woman $M=5.54$, $SD=.95$, $s^2=.91$).

There were no significant differences between the experimental conditions with regards to demographic characteristics (age, participant gender, race/ethnicity, and education),

indicating that random assignment was successful. We furthermore verified that the manipulations of user gender, $\chi^2(1, N=1,585)=1,220.48$, $p<.001$, and the complaint type, $\chi^2(4, N=1,583)=905.29$, $p<.001$, were successful.

Measures

The survey consisted of 23 questions and took an average completion time of 16 minutes. Variables were measured on 7-point scales from “*strongly disagree*” to “*strongly agree*,” unless indicated differently. Confirmatory factor analysis (CFA) was used to validate the structure of our data. In the following, we present means, standard deviations, and reliability measures for averaged items.

Outcome Variables: *Fairness* consisted of two concepts, procedural and outcome fairness. Studies revealed strong correlations, allowing to collapse both into one measurement (Gonçalves et al., 2023). Two items each, based on the study by Colquitt (2001) and Koper et al. (1993), measured perceived fairness of the moderation process (e.g., “The moderation of this post followed ethical and moral standards.”) and the moderation decision (e.g., “The removal of this post is justified.”; $M=5.64$; $SD=1.53$; $\alpha=.94$). For *legitimacy*, we adapted three items from the work of van der Toorn et al. (2011) to measure perceived legitimacy of the content removal (e.g., “This moderator is a legitimate authority and users should follow their decisions.”; $M=5.36$; $SD=1.4$; $\alpha=.85$). For *bias*, a two-item measurement assessed how much the moderator and the removal decision were perceived as impartial (e.g., “The moderation decision is biased.”). Using the Spearman-Brown formula for split-half reliability estimates (Eisinga et al., 2013) revealed a strong correlation between both items, which were thus collapsed into one measure ($M=3.13$; $SD=1.69$; Spearman-Brown coefficient=.94). In addition, we posed a simple yes/no question to check whether participants generally agreed with the removal of the post.

Mediating Variables: We operationalized perceived similarity as the degree to which participants were able to identify with the fictitious user through shared opinions and used a validated one-item measure of *identification* (“I identify with the social media user who made the post.”; Postmes et al., 2013) for this. Perceived similarities were further captured by three statements based on the study by Postmes et al. (2013) (e.g., “I think like the social media user who made the post”). The measurement demonstrated high reliability ($M=2.13$; $SD=1.52$; $\alpha=.94$). *Social influence* as a perceived characteristic of the fictitious user was measured with five statements adapted from a perceived influence scale (Salerno et al., 2019) (e.g., “The social media user who made the post is trustworthy.”; $M=1.87$; $SD=.88$; $\alpha=.87$).

Other Measures: To distinguish perceived similarity and identification from agreement with the hateful statement, we controlled for the latter with the single-item measure *agreement with the post* ($M=2.42$; $SD=1.8$). This was used as covariate in the analysis to account for participants agreeing

with the original statement and therefore being less favorable toward its removal.

We measured desirability of *gender-specific emotion traits* for women and men to gain insights into the prevalence of stereotypical gender perceptions in our sample by selecting 10 relevant traits (e.g., “aggressive,” “polite”) from the study by Prentice and Carranza (2002). We used an equal number of prescriptions or proscriptions that are intensified for one gender and relaxed for the other (e.g., prescription “assertive” relaxed for women and intensified for men, proscription “rebellious” intensified for women and relaxed for men).

Another relevant control variable was *concern for freedom of speech*. The three-item measurement focused on tolerance for the expression of opinions that deviate from one’s own and balancing freedom of speech with freedom from discrimination (Naab, 2012) (e.g., “Everyone has the right to express their opinion even if it differs from the majority”). After CFA, one reverse-coded item was excluded, leading to high internal consistency of the measurement ($M=5.29$; $SD=1.25$; Spearman-Brown coefficient = .82).

We controlled for *stance on immigration* because the stimulus contained an offensive statement against immigrants, by adapting four items from different representative polling questionnaires from Gallup and the Pew Research Center (Jones, 2019) (e.g., “American/Dutch/Portuguese identity, norms and values, are being threatened because there are too many immigrants in the U.S./Netherlands/Portugal.”; $M=4.55$; $SD=1.43$; $\alpha=.86$).

Results

Different mediation models were run with *complaint* and *user gender* as predictor variables, using the PROCESS macro for SPSS with 5,000 bootstrap samples. In the following, we will refer to higher degrees of perceived *fairness* and *legitimacy* and lower degrees of *bias* as positive evaluations of content moderation, while lower degrees of perceived *fairness* and *legitimacy* and higher degrees of *bias* indicate negative evaluations.

All reported coefficients are unstandardized and based on one composite sample ($n=1,586$; 51.7% identifying as women, mean age 46 years) from all three countries under study. An overview of the results separated by country is given in the Supplemental Material. *Country* was included as covariate in all analyses to account for possible cultural differences. We used indicator coding with the United States as reference category as most of the cited research on content moderation originates from a U.S. context.

Complaints About Moderation as a Predictor

In a serial mediation model, we tested how publicly contesting a content-removal decision affected evaluations of moderation. We used indicator coding for the categorical

predictor variable *complaint* (Hayes & Preacher, 2014) with “no complaint” as the reference category. The first mediator *identification* was expected to predict the second mediator *social influence* (Oldmeadow et al., 2003). *Agreement with the post*, *concern for freedom of speech*, *political leaning*, *stance on immigration*, and *country* were included as covariates, improving model fit for all outcome variables (without covariates $R^2_{\text{fairness}}=.157$, $R^2_{\text{legitimacy}}=.101$, $R^2_{\text{bias}}=.139$; with covariates $R^2_{\text{fairness}}=.262$, $R^2_{\text{legitimacy}}=.206$, $R^2_{\text{bias}}=.209$).

Our H1 predicted that a public complaint about a moderation decision influences participants to perceive moderation of a hateful post more negatively (less fair, less legitimate, more biased) than when no complaint is expressed. The complaints, however, had limited impacts on evaluations of moderation: In the free speech condition, participants did perceive moderation as significantly less fair than no complaint in the control group ($b=-.291$, $t=-3.37$, $p=.001$, CI $[-.461, -.122]$). However, the same complaint had no significant effect on perceptions of *legitimacy* or *bias* of moderation. The complaint alluding to political persecution did not have any significant effects on perceptions of moderation compared to the control group (see Table 1). H1, which assumed a negative relation between a public complaint about a content-removal decision and evaluation of moderation, is only partially supported for the free speech condition and *fairness* as the outcome variable.

The second hypothesis assumed that perceived similarity with the fictitious user (H2a) and perceived social influence of the fictitious user (H2b) would mediate effects of the complaint on evaluations of moderation. Analyses provided no evidence for mediation, as neither *complaint* condition predicted *identification* or *social influence*.⁵ Yet, both *identification* and *social influence* negatively affected evaluation of the moderation decision: Perceiving the fictitious user as similar to oneself and as having social influence led participants to make negative evaluations (see Table 1). In addition, *identification* predicted perceived *social influence* ($b=.339$, $t=23.042$, $p<.001$, CI $[.311, .368]$), confirming the assumption of serial mediation. The more participants identified with the fictitious user by perceiving him or her as similar to themselves, the more they perceived that user as socially influential (see Figure 1). While these results show that identification and social influence are highly important concepts in social media environments, both processes have limited impact for how moderation decisions are evaluated and seem unrelated to the complaints about a moderation decision.

User Gender as Predictor

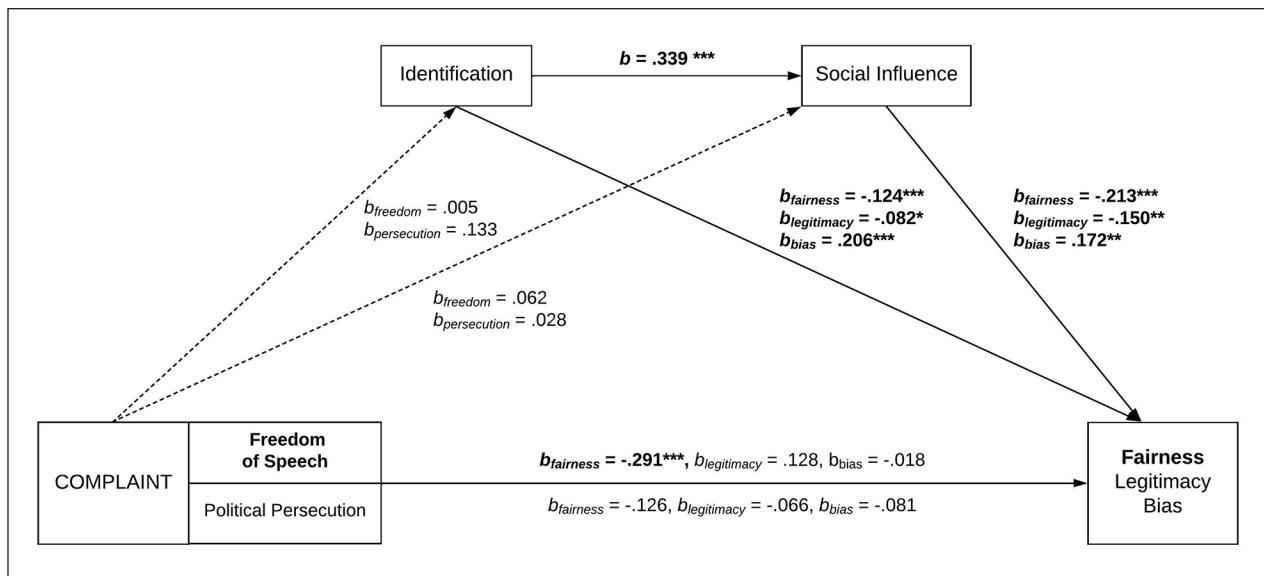
To test our hypotheses about gender and evaluations of moderation, we used a simple mediation model with *user gender* as the dichotomous predictor variable and *social influence* as the mediator. Including the covariates *agreement with the post*, *political leaning*, *stance on immigration*, and *country*

Table 1. Direct Effects of Mediation Model With Complaint as the Predictor and Identification and Social Influence as Mediators.

	Fairness		Legitimacy		Bias	
	<i>b</i> (SE)	95% CI	<i>b</i> (SE)	95% CI	<i>b</i> (SE)	95% CI
Complaint						
Free speech	-.291 (.09)***	[-.461, -.122]	-.128 (.08)	[-.288, .033]	-.018 (.10)	[-.209, .173]
Political persecution	-.126 (.09)	[-.295, .044]	-.066 (.08)	[-.226, .095]	-.081 (.10)	[-.273, .110]
Identification	-.124 (.04)***	[-.194, -.053]	-.082 (.03)*	[-.148, -.015]	.206 (.04)***	[.126, .285]
Social influence	-.213 (.05)***	[-.321, -.106]	-.150 (.05)**	[-.252, -.048]	.172 (.06)**	[.051, .293]
Covariates						
Agreement with post	-.032 (.03)	[-.083, .018]	-.011 (.02)	[-.058, .037]	.018 (.03)	[-.038, .075]
Concern for free speech	-.214 (.03)***	[-.272, -.157]	-.225 (.03)***	[-.279, -.171]	.154 (.03)***	[.09, .219]
Political leaning	-.048 (.02)**	[-.083, -.013]	-.040 (.02)*	[-.072, -.007]	.070 (.02)***	[.031, .109]
Stance on immigration	.244 (.03)***	[.186, .303]	.222 (.03)***	[.167, .277]	-.211 (.03)***	[-.277, -.145]
Country Netherlands	.522 (.09)***	[.346, .699]	.345 (.09)***	[.177, .512]	-.329 (.10)***	[-.529, -.13]
Country Portugal	.456 (.09)***	[.27, .641]	.199 (.09)*	[.024, .375]	.004 (.11)	[-.205, .214]

Note. Complaints referencing free speech limitations and political persecution arguments are compared to the reference category “no complaint.” For country, the Netherlands and Portugal are compared to the United States as reference category.

*Significant at $p \leq .05$. **Significant at $p \leq .01$. ***Significant at $p \leq .001$.

**Figure 1.** Mediation model with complaint as the predictor and identification and social influence as mediators.

*Significant at $p \leq .05$, **significant at $p \leq .01$, ***significant at $p \leq .001$. Direct effect of complaint on perception of moderation: coefficients in the upper row refer to the free speech complaint, and coefficients in the lower row refer to the political persecution complaint.

again improved the model fit for all outcome variables (without covariates $R^2_{\text{fairness}} = .096$, $R^2_{\text{legitimacy}} = .066$, $R^2_{\text{bias}} = .091$; with covariates $R^2_{\text{fairness}} = .263$, $R^2_{\text{legitimacy}} = .222$, $R^2_{\text{bias}} = .181$).

H3 predicted that when a hateful post by a woman is removed, participants would evaluate moderation more positively than when the removed post was made by a man. This assumption is not supported by our findings. *User gender* exerted no significant direct effects on perceptions of *fairness*, *legitimacy*, or *bias* of moderation (see Table 2). The key assumption behind H3 was that gender-specific emotion stereotypes are particularly violated when women act

aggressively, due to differences in the desirability of traits for women and men. To confirm, we conducted two exploratory factor analyses on the measured desirability of *gender-specific emotion traits* for women and men. We used principal components extraction with promax rotation based on Eigenvalues (>1.00). For women, the resulting model, $KMO = .74$, $\chi^2(N=1,556, 45) = 4319.32$, $p < .001$, had two factors and explained 52.7% of the variance in desirability, while for men, the model, $KMO = .78$, $\chi^2(N=1,565, 45) = 5524.31$, $p < .001$, had three factors and explained 66.6% of the variance.

Although the factor analyses uncovered different patterns, the items *aggressive*, *cynical*, *rebellious*, and *stubborn* were

Table 2. Direct Effects of Mediation Model With User Gender as the Predictor and Social Influence as the Mediator.

User gender	Social influence					
	<i>b</i> (SE) 95% CI					
	.116 (.04)** [.038, .195]					
	Fairness		Legitimacy		Bias	
	<i>b</i> (SE)	95% CI	<i>b</i> (SE)	95% CI	<i>b</i> (SE)	95% CI
Direct effects						
User gender	.041 (.07)	[-.102, .183]	.046 (.07)	[-.089, .181]	.097 (.08)	[-.062, .257]
Social influence	-.371 (.05)***	[-.465, -.277]	-.272 (.05)***	[-.361, -.184]	.366 (.05)***	[.261, .470]
	Effect (Boot SE)	Boot CI	Effect (Boot SE)	Boot CI	Effect (Boot SE)	Boot CI
Indirect effect	-.043 (.02)	[-.076, -.013]	-.032 (.01)	[-.058, -.009]	.043 (.02)	[.013, .077]
Covariates	<i>b</i> (SE)	95% CI	<i>b</i> (SE)	95% CI	<i>b</i> (SE)	95% CI
Agreement with post	-.065 (.02)**	[-.112, -.017]	-.032 (.02)	[-.077, .013]	.073 (.03)**	[.019, .126]
Political leaning	-.060 (.02)***	[-.096, -.025]	-.051 (.02)**	[-.084, -.017]	.083 (.02)***	[.043, .122]
Stance on immigration	.285 (.03)***	[.226, .343]	.257 (.03)***	[.201, .312]	-.259 (.03)***	[-.324, -.193]
Country Netherlands	.499 (.09)***	[.226, .343]	.321 (.09)***	[.149, .493]	-.293 (.10)**	[-.496, -.091]
Country Portugal	.444 (.10)***	[.257, .631]	.161 (.09)	[-.016, .338]	-.023 (.11)	[-.232, .186]

Note. For the predictor variable *user gender*, a woman is compared to the reference category of a man. For *country*, the Netherlands and Portugal are compared to the United States as a reference category.

*Significant at $p \leq .05$. **Significant at $p \leq .01$. ***Significant at $p \leq .001$. Significant indirect effects highlighted in gray.

loaded onto the same factor in both models. In theory, they are considered low in desirability, with the latter three being proscriptive for women (Prentice & Carranza, 2002). We grouped the items together into the variable *undesirable aggressive traits* for women ($M=3.48$; $SD=1.09$; $\alpha=.72$) and men ($M=3.50$; $SD=1.18$; $\alpha=.76$) and ran a paired samples *t*-test to determine differences. Results were not significant, $t(1,580)=-1.12$, $p=.264$, $CI[-.08, .02]$, and aggressive emotion traits were equally undesirable for both genders. Thus, there is no indication for the presumed double deviance of women acting aggressively by making a hateful post. This may help to explain why *user gender* had no direct effect on perceptions of moderation.

Finally, H4 hypothesized that the impact of the gender of the fictitious user on evaluations of content moderation is mediated by perceived social influence, presuming that a woman would be assigned less influence and moderation of her post would be evaluated more positively. Findings partially confirm this mediation effect, although the relation between *user gender*, *social influence*, and evaluations of moderation appears to be more complex. Again, perceived *social influence* of the fictitious user led to more negative evaluations of moderation (see Table 2). However, contrary to our expectations, *user gender* positively predicted *social influence* ($b=.116$, $t=2.91$, $p=.004$, $CI [.038, .195]$). Participants perceived the woman in our experiment as more socially influential than the man. There were significant indirect effects of *user gender* on all outcome variables, which were negative for *fairness* ($b=-.043$, $CI[-.076, -.013]$) and

legitimacy ($b=-.032$, $CI[-.058, -.009]$) and positive for *bias* ($b=.043$, $CI [.013, .077]$; see Figure 2). These results further illustrate the role of social influence in how other users are perceived and how removals of their content are evaluated. The gender of others represents an important cue based on which people make judgments about them, but perceptions of gender may be more complex and context-dependent than previously assumed.

Discussion

This study aimed to explain user perspectives on content moderation by disentangling how contextual factors such as arguments against moderation and interpersonal mechanisms evoked by perceived user characteristics influence evaluations of fairness, legitimacy, and bias of moderation. The findings complement existing knowledge on user views of social media governance, with implications for how content policies are executed and communicated.

Publicly complaining about a moderation decision influences how bystanders evaluate the fairness of content moderation. However, persuasiveness of such user complaints depends on the argument, and changes in how moderation was evaluated were driven by arguments about limitation to free speech. The fundamental nature of freedom of speech and its central role in discussions about moderation may trigger stronger reactions than a more ambiguous line of argumentation about political persecution of certain opinions (Van Noorloos, 2014). Even though the idea that social media

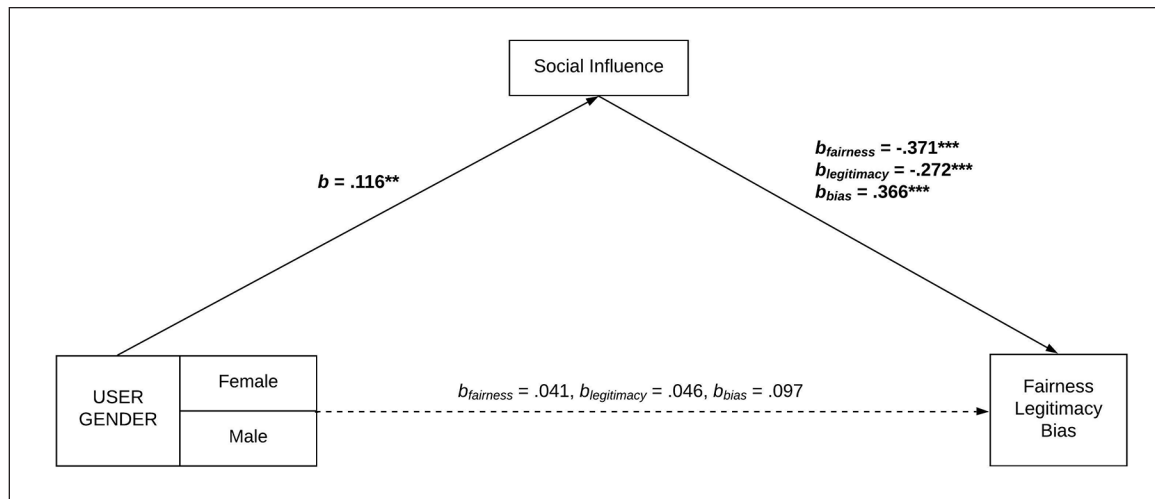


Figure 2. Mediation model with user gender as predictor.

*Significant at $p \leq .05$, **significant at $p \leq .01$, ***significant at $p \leq .001$.

platforms censor disagreeable political content is widespread among American conservatives (Vogels, 2022), this argument likely resonates less strongly with the general population⁶ than arguments about free speech, which are relatable for people with a diverse range of political opinions. Besides, the free speech argument only affected perceived fairness of moderation, but not legitimacy and bias. Participants may in principle be favorable toward content moderation but perceived it as rather unfair in the specific situation they witnessed in this experiment. Meanwhile general attitudes about legitimacy and bias of moderation may remain unaffected.

We found no evidence that participants evaluated the removal of a hateful post by a woman more positively, possibly because being aggressive was seen as undesirable for both men and women. This outcome stands in contrast to women being sanctioned for showing anger in other contexts (Brescoll & Uhlmann, 2008; Salerno & Peter-Hagene, 2015), as well as indications that perceptions of gender stereotypes are intensified in computer-mediated communication (Guegan et al., 2016). Another unexpected finding is the positive relation between another person's gender and their perceived social influence. Content moderation of a post clearly identifiable as hateful may constitute a case that overrides gender-related stereotypes, leading to positive evaluations of moderation regardless of the gender of the other user and associated stereotypical expectations. However, this point remains inconclusive, since other studies do not confirm the existence of such overriding effects (Wilhelm & Joeckel, 2019). Conclusions should furthermore be made with caution, as in certain contexts such as discussions of gender-based violence, women were found to have higher social influence (Carli, 2001). The hateful post in the experiment mentioned crime against women, possibly creating a context in which women hold more social influence than men regardless of emotion display.

While this study confirms that perceived similarity and social influence are relevant for how users evaluate moderation, both these mechanisms seem unrelated to the complaints. Our results do not allow for conclusions about the relevance of collective identities based on the arguments raised in the complaints. Nonetheless, a central takeaway from this study is that bystander perspectives on how harmful content is handled can indeed be influenced by cues about other users. Thus, not only do social media platforms apply some interpretative flexibility with their content policies—as examples of account bans and reinstatements show—but users as well are flexible in their evaluations of how content policies are enforced, and their judgments can be swayed by the perceived characteristics and argumentative strategies of other users.

From a theoretical point of view, the contribution of our study is twofold. First, we explored abstract ideas such as freedom of speech and political persecution as mechanisms of perceived similarity and identification. Previous studies on social identity (Hogg, 2021) used clearly defined ingroups and outgroups based on shared individual or collective traits to explore effects. We believed that, due to the nature of online discussions, abstract arguments could act as an identification or social influence mechanism, producing similar effects to more established groups. Given that the publicly posted complaints did not predict identification or social influence, our proposed expansion of social identity theories to include subscription to abstract ideals does not hold. However, a complaint about free speech limitations did have a significant effect on the perceived fairness of moderation. Our contribution is showing that social identity is an inadequate mechanism to explain this effect, allowing other scholars to explore alternative theoretical grounds for this, such as framing (Entman, 1993).

Our study also contributes to a theoretical understanding of social influence mechanisms. We tapped into the

long-standing strand of literature studying social influence based on gender and its effects (Eagly, 1983). Our findings, showing women being perceived as more influential than men, contradict many of the findings in this regard. However, as mentioned previously, we do not believe that this reflects an overall shift in perceptions of influence based on gender but, in contrast, highlights the contextuality of social influence. It shows how operationalizations of social influence, especially in low-information contexts such as online comment sections, need to be carefully contextualized. In this case, the interaction between the perceived gender identity of the commenter and the gender identity of the group mentioned in the comment is enough to overcome other social influence asymmetries. In fluid and multifaceted online settings, this means that previous work on social influence and gender should be revisited in light of its contextual settings.

Limitations

It should be critically noted that the experimental stimuli displayed no user comments or interaction metrics under the post, nor could participants obtain additional information, for example, through the fictitious user's profile. This limited how much information they had for discerning perceived similarity and social influence of the fictitious user. While these design choices exclude confounding influences, they may also limit the ecological validity of the findings.

We partly based H3 and H4 on the assumption that posting a racist, discriminatory statement equals the display of anger. Yet, group-directed hate as displayed here may also be associated with other emotions such as contempt and disgust (Martínez et al., 2022). Thus, the emotions attributed to the post and any gender-related stereotypical perceptions of the fictitious user may be more complex than assessed in this study. Besides, results may differ if the post contained a less blatantly hateful statement. A comparative study using a wider range of stimuli messages may provide clarification on this point.

In this study, we only considered evaluations of moderation of a White user breaching community guidelines, which disregards the influence of race or intersecting racial and gender identities. Future studies should include these aspects because stereotypical perceptions of anger and other (un)desirable forms of emotion expression are impacted by race (Salerno et al., 2019). There are known asymmetries in the interpretation of speech regulations, often prioritizing the protection of majority interests over the prevention of harms for the minority groups they are supposed to protect (Van Noorloos, 2014). Similar asymmetries most likely exist in the enforcement of online community guidelines, making race an important factor to include in related research. Such considerations should also be extended to approaching gender not as a binary concept but instead including nonbinary or genderfluid identities

into the research design. We acknowledge that our decision to focus on a binary interpretation of perceived gender risks reinforcing a binary construction of gender in research design and see this as a key ethical limitation of our study.

Implications

In line with the study by Riedl et al. (2022), we argue that both the scientific community and decision-makers in content moderation benefit from a more comprehensive understanding of user perspectives on moderation. User interactions and, in a broader sense, public debates are shaped by content guidelines. But platforms also should have an interest in implementing fair and appropriate moderation practices to not lose users to competitors on an increasingly diverse spectrum of social media platforms.

The fact that the free speech complaint resonated with our participants and led them to evaluate moderation as less fair emphasizes the considerable impact of this argument in the discourse about content moderation. This is also evident in how fears about free speech limitations have been intentionally politicized to shape public perceptions of hate speech regulation (Van Noorloos, 2014). Such allegations should be taken seriously by social media platforms as they could harm the perception of an important practice like content moderation. Findings on baseline effects from experimental studies like this can provide food for thought for different social media platforms for devising strategies that emphasize the broader societal benefits of content moderation to their users while diffusing false claims about limitations to free speech. The results of this research may also encourage critical discussions of content policies and holding platforms publicly accountable for their moderation practices. While paying attention to existing asymmetries, these discussions should not be co-opted by unsubstantiated accusations of moderation lacking fairness or legitimacy. Thereby, attention should also be paid to how mechanisms related to social influence, perceptions of stereotypical user characteristics, and identities can give prominent counterarguments against moderation more weight in certain contexts. To avoid allegations of censorship and foster trust with a well-informed user base, platforms are advised to be more transparent and sufficiently explain how they enforce content guidelines.

Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: This study was made possible through a research award by Meta. The topic of the study relates to Meta's business activities, creating a potential conflict of interest. However, Meta was not involved at any stage of the design, execution, or reporting of this study. Funds were awarded as an unrestricted gift, and all research was carried out independently by the authors without any external interference or influence.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The authors received financial support from the Content Policy Research on Social Media Platforms Award by Meta (then Facebook) in 2019.

ORCID iDs

Ina Weber  <https://orcid.org/0000-0002-3501-6905>

Marisa Torres da Silva  <https://orcid.org/0000-0003-1136-4232>

Joep Hofhuis  <https://orcid.org/0000-0001-7531-8644>

Supplemental Material

Supplemental material for this article is available online.

Notes

1. It should be noted that the data were collected during the first months of the COVID-19 pandemic as well as during a moment when the United States was having significant and contested elections. The current events at the time were also reflected in conversations around content moderation (e.g., in relation to the spread of misinformation about COVID-19 on social media and Twitter's actions against tweets by then-president Trump that could promote violent actions).
2. In the United States, commonly used categories for race/ethnicity were applied. Since these do not translate well to the European contexts, they were adjusted to questions about the participants' and their parents' country of birth in the Netherlands. No questions concerning race/ethnicity were asked to participants in Portugal due to constitutional regulations that forbid to do so. Quotas for sampling were retrieved from Pew Research Center for the United States, from the national statistics institutes of the Netherlands (CBS Centraal Bureau voor de Statistiek) and Portugal (INE Instituto Nacional de Estatística), as well as Eurostat.
3. In total, 1,976 participants filled in the survey. From this number, we excluded participants who skipped ($n=9$) or failed ($n=345$) the attention check and who completed the survey faster than 48% of the median survey duration ($n=45$), resulting in a sample of 1,586 participants ($n=533$ in the United States, $n=530$ in the Netherlands, and $n=523$ in Portugal).
4. cf. <https://generated.photos/faces> with the following parameters: natural, adult, White, brown hair, neutral expression.
5. For *identification*: Free speech $b=-.005$, $t=-.062$, $p=.951$, CI $[-.150, .141]$; political persecution $b=.133$, $t=1.790$, $p=.074$, CI $[-.013, .278]$. For *social influence*: Free speech $b=.062$, $t=-1.482$, $p=.139$, CI $[-.020, .143]$; political persecution $b=.028$, $t=.663$, $p=.508$, CI $[-.054, .109]$.
6. It should be noted that political leaning was normally distributed in the sample; hence, there are no higher proportions of left-leaning participants inclined to reject this argument or right-leaning participants inclined to accept it. Furthermore, separate analyses show no differences in the absence of effects among the United States, the Netherlands, and Portugal, ruling out the possibility that the political persecution argument resonated only in certain country contexts.

References

- Abrams, D., Travaglino, G. A., Marques, J. M., Pinto, I., & Levine, J. M. (2018). Deviance credit: Tolerance of deviant ingroup leaders is mediated by their accrual of prototypicality and conferral of their right to be supported. *Journal of Social Issues*, 74(1), 36–55. <https://doi.org/10.1111/josi.12255>
- Althaus, S. L., & Coe, K. (2011). Priming patriots: Social identity processes and the dynamics of public support for war. *Public Opinion Quarterly*, 75(1), 65–88. <https://doi.org/10.1093/poq/nfq071>
- Biernat, M., & Sesko, A. K. (2018). Gender stereotypes and stereotyping: A cognitive perspective on gender bias. In N. K. Dess, J. Marecek, & L. C. Bell (Eds.), *Gender, sex, and sexualities: Psychological perspectives* (pp. 171–194). Oxford University Press.
- Bleich, E. (2011). The rise of hate speech and hate crime laws in liberal democracies. *Journal of Ethnic and Migration Studies*, 37(6), 917–934. <https://doi.org/10.1080/1369183X.2011.576195>
- Brescoll, V. L., & Uhlmann, E. L. (2008). Can an angry woman get ahead? Status conferral, gender, and expression of emotion in the workplace. *Psychological Science*, 19(3), 268–275. <https://doi.org/10.1111/j.1467-9280.2008.02079.x>
- Burgoon, M., Pauls, V., & Roberts, D. L. (2002). Language expectancy theory. In J. Dillard & M. Pfau (Eds.), *The persuasion handbook: Developments in theory and practice* (pp. 117–136). Sage. <https://doi.org/10.4135/9781412976046.n7>
- Carli, L. L. (2001). Gender and social influence. *Journal of Social Issues*, 57(4), 725–741. <https://doi.org/10.1111/0022-4537.00238>
- Celuch, M., Oksanen, A., Räsänen, P., Costello, M., Blaya, C., Zych, I., Llorent, V. J., Reichelmann, A., & Hawdon, J. (2022). Factors associated with online hate acceptance: A cross-national six-country study among young adults. *International Journal of Environmental Research and Public Health*, 19(1), 534. <https://doi.org/10.3390/ijerph19010534>
- Colquitt, J. A. (2001). On the dimensionality of organizational justice: A construct validation of a measure. *Journal of Applied Psychology*, 86(3), 386–400. <https://doi.org/10.1037/0021-9010.86.3.386>
- Colquitt, J. A., Greenberg, J., & Zapata-Phelan, C. P. (2013). What is organizational justice? A historical overview. In J. Greenberg & J. A. Colquitt (Eds.), *Handbook of organizational justice* (pp. 3–56). Psychology Press.
- Craciun, G., & Moore, K. (2019). Credibility of negative online product reviews: Reviewer gender, reputation and emotion effects. *Computers in Human Behavior*, 97, 104–115. <https://doi.org/10.1016/j.chb.2019.03.010>
- Croes, E., & Bartels, J. (2021). Young adults' motivations for following social influencers and their relationship to identification and buying behavior. *Computers in Human Behavior*, 124, 106910. <https://doi.org/10.1016/j.chb.2021.106910>
- Eagly, A. H. (1983). Gender and social influence: A social psychological analysis. *American Psychologist*, 38(9), 971–981. <https://doi.org/10.1037/0003-066X.38.9.971>
- Eisinga, R., Grotenhuis, M. T., & Pelzer, B. (2013). The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? *International Journal of Public Health*, 58(4), 637–642. <https://doi.org/10.1007/s00038-012-0416-3>

- Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4), 51–58. <https://doi.org/10.1111/j.1460-2466.1993.tb01304.x>
- European Commission. (2017). *Gender equality 2017: Gender equality, stereotypes, and women in politics* (465; Special Eurobarometer). European Commission, Directorate-General for Communication. https://data.europa.eu/data/datasets/s2154_87_4_465_eng?locale=en
- Gerrard, Y., & Thornham, H. (2020). Content moderation: Social media's sexist assemblages. *New Media & Society*, 22(7), 1266–1286. <https://doi.org/10.1177/1461444820912540>
- Gnambs, T., & Batinić, B. (2013). The roots of interpersonal influence: A mediated moderation model for knowledge and traits as predictors of opinion leadership. *Applied Psychology*, 62(4), 597–618. <https://doi.org/10.1111/j.1464-0597.2012.00497.x>
- Gonçalves, J., Weber, I., Masullo, G. M., Torres Da Silva, M., & Hofhuis, J. (2023). Common sense or censorship: How algorithmic moderators and message type influence perceptions of online content deletion. *New Media & Society*, 25(10), 2595–2617. <https://doi.org/10.1177/14614448211032310>
- Guegan, J., Moliner, P., & Milland, L. (2016). Social asymmetries and anonymity in dyadic computer-mediated communication: An experimental study of gender perception. *Swiss Journal of Psychology*, 75(1), 15–23. <https://doi.org/10.1024/1421-0185/a000167>
- Gupta, P., Burton, J. L., & Costa Barros, L. (2023). Gender of the online influencer and follower: The differential persuasive impact of homophily, attractiveness and product-match. *Internet Research*, 33(2), 720–740. <https://doi.org/10.1108/INTR-04-2021-0229>
- Haimson, O. L., Delmonaco, D., Nie, P., & Wegner, A. (2021). Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2), 1–35. <https://doi.org/10.1145/3479610>
- Hasler, B. S., & Amichai-Hamburger, Y. (2013). Online intergroup contact. In Y. Amichai-Hamburger (Ed.), *The social net* (pp. 220–252). Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199639540.003.0012>
- Hayes, A. F., & Preacher, K. J. (2014). Statistical mediation analysis with a multicategorical independent variable. *British Journal of Mathematical and Statistical Psychology*, 67(3), 451–470. <https://doi.org/10.1111/bmsp.12028>
- Heller, B., & van Hoboken, J. (2019). *Freedom of expression: A comparative summary of United States and European Law* (Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression) [Working Paper]. Institute for Information Law, University of Amsterdam. https://www.ivir.nl/publicaties/download/TWG_Freedom_of_Expression.pdf
- Hogg, M. A. (2005). Uncertainty, social identity, and ideology. In S. R. Thye & E. J. Lawler (Eds.), *Advances in group processes* (Vol. 22, pp. 203–229). Emerald. [https://doi.org/10.1016/S0882-6145\(05\)22008-8](https://doi.org/10.1016/S0882-6145(05)22008-8)
- Hogg, M. A. (2021). Self-uncertainty and group identification: Consequences for social identity, group behavior, intergroup relations, and society. In B. Gawronski (Ed.), *Advances in experimental social psychology* (Vol. 64, pp. 263–316). Academic Press. <https://doi.org/10.1016/bs.aesp.2021.04.004>
- Hurwitz, J., & Mondak, J. J. (2002). Democratic principles, discrimination and political intolerance. *British Journal of Political Science*, 32(1), 93–118. <https://doi.org/10.1017/S0007123402000042>
- Jhaver, S., Appling, D. S., Gilbert, E., & Bruckman, A. (2019). “Did you Suspect the post would be removed?” Understanding user reactions to content removals on Reddit. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 1–33. <https://doi.org/10.1145/3359294>
- Jhaver, S., Bruckman, A., & Gilbert, E. (2019). Does transparency in moderation really matter? User behavior after content removal explanations on Reddit. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 1–27. <https://doi.org/10.1145/3359252>
- Jones, B. (2019). *Majority of Americans continue to say immigrants strengthen the U.S.* Pew Research Center. https://www.pewresearch.org/wp-content/uploads/2019/01/FT_19.01.29_ImmigrationBurden_Topline.pdf
- Koper, G., Van Knippenberg, D., Bouhuijs, F., Vermunt, R., & Wilke, H. (1993). Procedural fairness and self-esteem. *European Journal of Social Psychology*, 23(3), 313–325. <https://doi.org/10.1002/ejsp.2420230307>
- Mac, R., & Browning, K. (2022, November 19). Elon Musk reinstates Trump's Twitter account. *The New York Times*. <https://www.nytimes.com/2022/11/19/technology/trump-twitter-musk.html>
- Martínez, C. A., van Prooijen, J.-W., & Van Lange, P. A. M. (2022). Hate: Toward understanding its distinctive features across interpersonal and intergroup targets. *Emotion*, 22(1), 46–63. <https://doi.org/10.1037/emo0001056>
- Melamed, D., & Savage, S. V. (2016). Status, faction sizes, and social influence: Testing the theoretical mechanism. *American Journal of Sociology*, 122(1), 201–232. <https://doi.org/10.1086/686943>
- Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, 20(11), 4366–4383. <https://doi.org/10.1177/1461444818773059>
- Naab, T. (2012). The relevance of people's attitudes towards freedom of expression in a changing media environment. *ESSACHESS Journal for Communication Studies*, 5(1), 45–67. <https://ssrn.com/abstract=2161607>
- Oldmeadow, J. A., Platow, M. J., Foddy, M., & Anderson, D. (2003). Self-categorization, status, and social influence. *Social Psychology Quarterly*, 66(2), 138. <https://doi.org/10.2307/1519844>
- Park, K., Ging, D., Murphy, S., & McGrath, C. (2023). *The impact of the use of social media on women and girls* [Policy Department for Citizens' Rights and Constitutional Affairs]. European Parliament. [https://www.europarl.europa.eu/think-tank/en/document/IPOL_STU\(2023\)743341](https://www.europarl.europa.eu/think-tank/en/document/IPOL_STU(2023)743341)
- Pettersson, K. (2019). “Freedom of speech requires actions”: Exploring the discourse of politicians convicted of hate-speech against Muslims. *European Journal of Social Psychology*, 49, 938–952. <https://doi.org/10.1002/ejsp.2577>
- Pew Research Center. (2017). *On gender differences, no consensus on nature vs. nurture*. <https://www.pewresearch.org/social-trends/2017/12/05/on-gender-differences-no-consensus-on-nature-vs-nurture/>
- Postmes, T., Haslam, S. A., & Jans, L. (2013). A single-item measure of social identification: Reliability, validity, and utility.

- British Journal of Social Psychology*, 52(4), 597–617. <https://doi.org/10.1111/bjso.12006>
- Postmes, T., Haslam, S. A., & Swaab, R. I. (2005). Social influence in small groups: An interactive model of social identity formation. *European Review of Social Psychology*, 16(1), 1–42. <https://doi.org/10.1080/104632804400000062>
- Prentice, D. A., & Carranza, E. (2002). What women and men should be, shouldn't be, are allowed to be, and don't have to be: The contents of prescriptive gender stereotypes. *Psychology of Women Quarterly*, 26(4), 269–281. <https://doi.org/10.1111/1471-6402.t01-1-00066>
- Rashotte, L. (2007). Social influence. In G. Ritzer (Ed.), *The Blackwell encyclopedia of sociology* (pp. 4434–4437). John Wiley & Sons, Ltd. <https://doi.org/10.1002/9781405165518.wbeoss154>
- Rathje, S., Van Bavel, J. J., & Van Der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, 118(26), Article e2024292118. <https://doi.org/10.1073/pnas.2024292118>
- Raven, B. H. (2008). The bases of power and the power/interaction model of interpersonal influence. *Analyses of Social Issues and Public Policy*, 8(1), 1–22. <https://doi.org/10.1111/j.1530-2415.2008.00159.x>
- Riedl, M. J., Whipple, K. N., & Wallace, R. (2022). Antecedents of support for social media content moderation and platform regulation: The role of presumed effects on self and others. *Information, Communication & Society*, 25(11), 1632–1649. <https://doi.org/10.1080/1369118X.2021.1874040>
- Robertson, A. (2022, April 15). What Elon Musk's Twitter "free speech" promises miss. *The Verge*. <https://www.theverge.com/2022/4/15/23025120/elon-musk-twitter-free-speech-government-censorship>
- Rothenberg, E. (2023, July 30). Elon Musk reactivates Kanye West's Twitter account following X rebrand. *CNN*. <https://edition.cnn.com/2023/07/30/business/kanye-west-twitter-account-elon-musk/index.html>
- Salerno, J. M., & Peter-Hagene, L. C. (2015). One angry woman: Anger expression increases influence for men, but decreases influence for women, during group deliberation. *Law and Human Behavior*, 39(6), 581–592. <https://doi.org/10.1037/lhb0000147>
- Salerno, J. M., Peter-Hagene, L. C., & Jay, A. C. V. (2019). Women and African Americans are less influential when they express anger during group decision making. *Group Processes & Intergroup Relations*, 22(1), 57–79. <https://doi.org/10.1177/1368430217702967>
- Shaughnessy, B., DuBosar, E., Hutchens, M. J., & Mann, I. (2024). An attack on free speech? Examining content moderation, (de-) and (re-) platforming on American right-wing alternative social media. *New Media & Society*. Advance online publication. <https://doi.org/10.1177/14614448241228850>
- Stollznow, K. (2020). *On the offensive: Prejudice in language past and present* (1st ed.). Cambridge University Press. <https://doi.org/10.1017/9781108866637>
- Suzor, N., Van Geelen, T., & Myers West, S. (2018). Evaluating the legitimacy of platform governance: A review of research and a shared research agenda. *International Communication Gazette*, 80(4), 385–400. <https://doi.org/10.1177/1748048518757142>
- Suzor, N., West, S., Quodling, A., & York, J. (2019). What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation. *International Journal of Communication*, 13(18), 1526–1543. <https://ijoc.org/index.php/ijoc/article/view/9736>
- Tajfel, H., Turner, J. C., Austin, W. G., & Worchel, S. (1979). An integrative theory of intergroup conflict. In M. J. Hatch & M. Schultz (Eds.), *Organizational identity: A reader* (pp. 56–65). Oxford University Press.
- Trevelyan, M. (2022, March 11). Facebook owner defends policy on calls for violence that angered Russia. *Reuters*. <https://www.reuters.com/article/ukraine-crisis-meta-russia-idCNL5N2VE2TP>
- United Nations Development Programme. (2024). *Human development report 2023-24: Breaking the gridlock: Reimagining cooperation in a polarized world*. <https://hdr.undp.org/content/human-development-report-2023-24>
- Vaccaro, K., Sandvig, C., & Karahalios, K. (2020). "At the end of the day Facebook does what it wants": How users experience contesting algorithmic content moderation. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2), 1–22. <https://doi.org/10.1145/3415238>
- Vaidhyanathan, S. (2019, July 28). Why conservatives allege Big Tech is muzzling them. *The Atlantic*. <https://www.theatlantic.com/ideas/archive/2019/07/conservatives-pretend-big-tech-biased-against-them/594916/>
- Van Bavel, J. J., & Packer, D. J. (2021). *The power of us: Harnessing our shared identities to improve performance, increase cooperation, and promote social harmony* (1st ed.). Little, Brown Spark.
- van der Toorn, J., Tyler, T. R., & Jost, J. T. (2011). More than fair: Outcome dependence, system justification, and the perceived legitimacy of authority figures. *Journal of Experimental Social Psychology*, 47(1), 127–138. <https://doi.org/10.1016/j.jesp.2010.09.003>
- van Dijke, M., De Cremer, D., & Mayer, D. M. (2010). The role of authority power in explaining procedural fairness effects. *Journal of Applied Psychology*, 95(3), 488–502. <https://doi.org/10.1037/a0018921>
- Van Noorloos, M. (2014). The politicisation of hate speech bans in the twenty-first-century Netherlands: Law in a changing context. *Journal of Ethnic and Migration Studies*, 40(2), 249–265. <https://doi.org/10.1080/1369183X.2013.851474>
- Vogels, E. A. (2022). *Support for more regulation of tech companies has declined in U.S., especially among Republicans*. Pew Research Center. <https://www.pewresearch.org/short-reads/2022/05/13/support-for-more-regulation-of-tech-companies-has-declined-in-u-s-especially-among-republicans/>
- Vogels, E. A., Perrin, A., & Anderson, M. (2020). *Most Americans think social media sites censor political viewpoints*. Pew Research Center. <https://www.pewresearch.org/internet/2020/08/19/most-americans-think-social-media-sites-censor-political-viewpoints/>
- Wagner, C., Garcia, D., Jadidi, M., & Strohmaier, M. (2015). *It's a Man's Wikipedia? Assessing gender inequality in an online encyclopedia*. <https://doi.org/10.48550/ARXIV.1501.06307>
- Walther, J. B., Carr, C. T., & Choi, S. S. W. (2011). Interaction of interpersonal, peer, and media influence sources online: A research agenda for technology convergence. In Z. Papacharissi (Ed.), *A networked self: Identity, community and culture on social network sites* (pp. 25–46). Routledge.
- Wang, S. S., Moon, S.-I., Kwon, K. H., Evans, C. A., & Stefanone, M. A. (2010). Face off: Implications of visual cues on initiating

friendship on Facebook. *Computers in Human Behavior*, 26(2), 226–234. <https://doi.org/10.1016/j.chb.2009.10.001>

Weber, I., Laban, A., Masullo, G. M., Gonçalves, J., Torres Da Silva, M., & Hofhuis, J. (2020). *International perspectives on what's considered hateful or profane online* [White Paper]. Center for Media Engagement. <https://mediaengagement.org/wp-content/uploads/2020/09/CME-Report-International-Perspectives-on-What's-Considered-Hateful-or-Profane-Online-1.pdf>

West, E. A., & Iyengar, S. (2022). Partisanship as a social identity: Implications for polarization. *Political Behavior*, 44(2), 807–838. <https://doi.org/10.1007/s11109-020-09637-y>

Wilhelm, C. (2020). Gendered (in)visibility in digital media contexts. *Studies in Communication Sciences*. <https://doi.org/10.24434/j.scoms.2021.01.007>

Wilhelm, C., & Joeckel, S. (2019). Gendered morality and backlash effects in online discussions: An experimental study on how users respond to hate speech comments against women and sexual minorities. *Sex Roles*, 80(7–8), 381–392. <https://doi.org/10.1007/s11199-018-0941-5>

Author Biographies

Ina Weber (M.A. Erasmus University Rotterdam) is a PhD student of Communication Studies at the University of Antwerp. Her

research interests include online hate speech, content moderation, and technological interventions.

João Gonçalves (Ph.D. Universidade do Minho) is Associate Professor of Media and Communication Studies at Erasmus University Rotterdam. His research interests include political communication and artificial intelligence.

Gina M. Masullo (Ph.D. Syracuse University) is Associate Professor in the School of Journalism and Media at the University of Texas at Austin. Her research interests include online incivility, political participation, and connective democracy.

Marisa Torres da Silva (Ph.D. Universidade Nova de Lisboa) is Associate Professor in the Department of Communication Sciences at Universidade Nova de Lisboa and researcher at Instituto de Comunicação da Nova (ICNOVA). Her research interests include public sphere, gender, and communication and online hate speech.

Joep Hofhuis (Ph.D. University of Groningen) is an associate professor at the Amsterdam School of International Business of Amsterdam University of Applied Sciences. His research interests include intercultural communication and workplace diversity.

Appendix

Examples of the Stimuli



Figure A1. Hateful social media post by fictitious user, example for user gender woman.



Figure A2. Hateful social media post by fictitious user, example for user gender man.

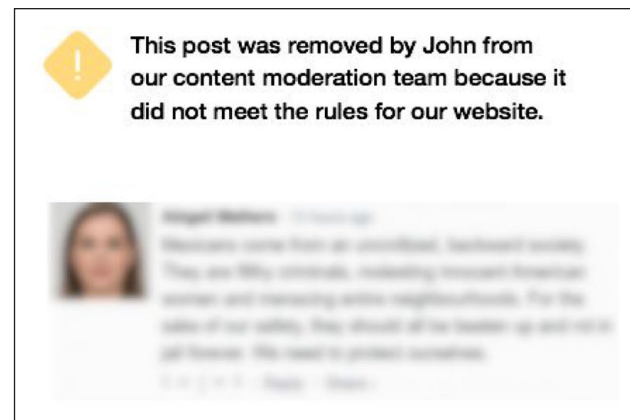


Figure A3. Moderation message informing about the removal of the post.



Figure A4. Complaint about removal of the post, argument about freedom of speech.