

Breaking the Silence: Investigating Which Types of Moderation Reduce Negative Effects of Sexist Social Media Content

JULIA SASSE*, Ansbach University of Applied Sciences, Germany

JENS GROSSKLAGS, Technical University of Munich, Germany

Sexist content is widespread on social media and can reduce women's psychological well-being and their willingness to participate in online discourse, making it a societal issue. To counter these effects, social media platforms employ moderators. To date, little is known about the effectiveness of different forms of moderation in creating a safe space and their acceptance, in particular from the perspective of women as members of the targeted group and users in general (rather than perpetrators). In this research, we propose that some common forms of moderation can be systematized along two facets of visibility, namely visibility of sexist content and of counterspeech. In an online experiment ($N = 839$), we manipulated these two facets and tested how they shaped social norms, feelings of safety, and intent to participate, as well as fairness, trustworthiness, and efficacy evaluations. In line with our predictions, deletion of sexist content – i.e., its invisibility – and (public) counterspeech – i.e., its visibility – against visible sexist content contributed to creating a safe space. Looking at the underlying psychological mechanism, we found that these effects were largely driven by changes in what was perceived normative in the presented context. Interestingly, deletion of sexist content was judged as less fair than counterspeech against visible sexist content. Our research contributes to a growing body of literature that highlights the importance of norms in creating safer online environments and provides practical implications for moderators for selecting actions that can be effective and accepted.

CCS Concepts: • **Human-centered computing** → **Social networking sites**; **User studies**; **Human computer interaction (HCI)**; **Collaborative and social computing**.

Additional Key Words and Phrases: social networking site design and use, behavior change, computer mediated communication, gender and identity, social media and online communities, quantitative methods

ACM Reference Format:

Julia Sasse and Jens Grossklags. 2023. Breaking the Silence: Investigating Which Types of Moderation Reduce Negative Effects of Sexist Social Media Content. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 327 (October 2023), 26 pages. <https://doi.org/10.1145/3610176>

1 INTRODUCTION

Sexual harassment and sexist insults are frequent in social media, and in most cases, young women are the target [2, 19, 55]. As a consequence, targeted women may experience stress, anxiety, panic attacks, or lowered self-esteem [2] and frequently change their online behavior, up to the point of withdrawing from conversations and turning silent [33]. Reducing the impact of such attacks on women (as well as other forms of misconduct) is one of the main tasks of platform moderation. Critically, moderators can execute this task in different ways, and it is thus far under-researched

*Research conducted while affiliated with the Technical University of Munich, Germany.

Authors' addresses: Julia Sasse, julia.sasse@hs-ansbach.de, Ansbach University of Applied Sciences, Residenzstrasse 8, Ansbach, Germany; Jens Grossklags, Technical University of Munich, Boltzmannstr. 3, Garching, Germany, jens.grossklags@in.tum.de.



This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike International 4.0 License](https://creativecommons.org/licenses/by-nc-sa/4.0/).

© 2023 Copyright held by the owner/author(s).

2573-0142/2023/10-ART327

<https://doi.org/10.1145/3610176>

which forms of moderation are well accepted and effective in creating a space where women feel safe and motivated to speak up.

In the here presented research, we investigated the extent to which facets of *moderation visibility*, namely visibility of initial sexist attacks and visibility of their reprimand, determine the effectiveness and acceptance of moderation. We did so by considering both the perspective of female social media users as members of the targeted group as well as of men to obtain a comprehensive understanding of beneficial and potential adverse effects. Our approach brings social media users to the foreground and thus complements recent work on moderation that often estimates effectiveness based on perpetrators' reactions [e.g., 21, 22, 24, 51] or the degree of automation sophistication [e.g., 30, 41, 56], thereby providing novel user-based insights for the governance of social media platforms, for instance for the training of moderators and the design of supporting tools.

1.1 Sexism Online

In general terms, sexism refers to prejudice against or discrimination of a person or a group of people merely based on their sex or gender by individuals, groups, or institutions [49]. Accordingly, individuals of all sexes and genders may be subject to sexist treatment; yet, in general, women experience sexism more often than men [37, 50]. While sexism is at times unambiguous and easy to detect (e.g., sexual harassment), it may also take more subtle forms (e.g., putting female job candidates up for unfair competition), and it can be hostile (e.g., sexist insults) or benevolent (e.g., overprotection of women) [14, 47]. In its essence, no matter the form, sexism contributes to the maintenance of gender-based inequality with regard to power and status in society [15].

Since the internet has developed from an information platform into a space for active participation and social exchange, it has also broadened the scope of how women as a group are disparaged, for instance, through the use of hate speech, and how individuals are sexually approached, objectified, or harassed [19]. A 2017 survey by Amnesty International shows that such events are frequent: In eight countries, from Poland to New Zealand, 23% of women reported that they had experienced online abuse or harassment [2]. A survey on online sexual harassment and cyberstalking among young women living in the European Union, commissioned by the European Parliament's Policy Department for Citizens' Rights and Constitutional Affairs, produced similar results [53]. Moreover, survey data from the United States show a critical trend: Between the years 2017 and 2020, the percentage of women who reported having experienced sexual harassment online had doubled and was three times higher than for men [55].

Sexist online behavior has substantial negative implications for individuals and society as a whole: On the one hand, it negatively impacts self-esteem and mental health, and on the other hand, it reduces women's participation in the online discourse, posing obstacles to gender equality [2, 20, 53, 54]. As targets, women are often not willing to speak up against sexism [43, 48], nor should it be their responsibility. Against this backdrop, it is pivotal that online platforms take action against sexist behavior shown by their users.

1.2 Content Moderation

Identifying and regulating sexist behavior – as well as other forms of harmful conduct – falls within the purview of moderators who may take on the task on a voluntary or professional basis [26, 45]. On most platforms, moderators can choose from a range of different actions in response to harmful conduct [23]. According to an interview study by Seering and colleagues [45], moderators often respond first through verbal warnings, deleting content, and finally by (temporarily) banning offenders; at times, they also explain to offenders their moderation actions. Noteworthy, while all platforms provide some guidelines that users shall abide by, regulations and their enforcement are often fuzzy [38], presumably making it often difficult for moderators to (consistently) judge

and select appropriate actions, which often constitute opposite choices [23]. Notwithstanding these difficulties, it is pivotal that moderation takes place since harmful content in an unrestricted environment leads to more harmful content and thus deteriorates the discourse [29].

Critically, different types of moderation may produce different effects. For instance, Jhaver and colleagues [21, 22] investigated how offenders responded to content deletion on Reddit. Deletion was perceived as fairer and associated with fewer future content deletions if combined with an explanation, suggesting that transparency may facilitate positive effects of moderation, at least from the perspective of perpetrators; an interpretation that is supported by similar findings by Tyler and colleagues [51]. Also, targets of sexual harassment appear to have preferences for certain types of moderation. Im and colleagues [20] conducted a scenario study and found that women preferred the removal of content in which they were insulted or disrespected; they were also supportive of labeling harassing content as such and of banning perpetrators. Conversely, revealing the identity of perpetrators or making them pay reparations were less preferred options. While these findings do not provide direct insight into the effectiveness of different forms of moderation, the fact that women are not indifferent to the question of how sexist content is handled suggests that there may well be differences. Such findings are evidence that it is necessary to carefully and systematically differentiate different forms of content moderation and to stake out their consequences in order to create a less hostile environment.

We argue that the concept of *visibility* serves well to systematize some of the most frequently used forms of moderation. Visibility may apply to different aspects of moderation; here, we focus on the visibility of the offense (i.e., sexism) and of its reprimand (i.e., counterspeech). First, in case of deletion, the offense is no longer visible, while in case of many other forms of moderation (such as warnings) or inaction, the offense remains visible. Second, moderation may differ in the extent to which an offense is publicly (and hence visibly) reprimanded. Such reprimand by moderators constitutes a form of *counterspeech* which refers to any form of “communication that seeks to counteract potential harm that is brought about by other speech” [7, p. 2]. From the definition, it is apparent that counterspeech could be voiced by any target or witness of harmful content [7, 30]. Since we focus our investigation on counterspeech as a form of content moderation, in the present context, we only consider counterspeech from moderators. A moderator may engage in public counterspeech but may also choose to do so in a private message to the offender, or not at all. While private counterspeech reveals a moderator’s judgment and motives to the perpetrator alone, public counterspeech does so to the wider community. As apparent in the research by Jhaver and colleagues [21, 22], the two facets of visibility may be combined in various constellations. In the present research, we tested six constellations of sexism visibility and counterspeech visibility (see Fig. 1 for an example).

Previous research has looked at other facets of visibility in the context of moderation, in particular, with regard to the source of moderation [5, 35]. In a realistic social media setting, Bhandari and colleagues [5] varied the extent to which information about a moderator was visible and found that this affected subsequent user interventions (i.e., flagging through bystanders) of unmoderated harassing posts. If information about the source of the moderation (i.e., another user or an automated system) was visible, fewer user-based interventions occurred than when the source of moderation was unknown. While this research investigated a different facet of visibility, it supports the notion that visibility may be a critical factor for the effectiveness of moderation.

We expect that sexism and counterspeech visibility may be critical for determining the effectiveness of different forms of moderation with regard to social media users in general and in particular for members of the targeted group. This is because visible expressions of attitudes and opinions may serve as *normative signals* to others since they provide an indication of which views and actions are prevalent and tolerated in a given environment [9]. Against this theoretical backdrop,

we predicted that different constellations of visibility lead to different normative signals and shape the extent to which in particular female users feel safe.

1.3 The Role of Social Norms

Social norms are derived from how individuals behave or how they are expected to behave in a given environment. These two types of norms are referred to as descriptive (i.e., describing actual behavior) and injunctive (i.e., stating what is approved or disapproved behavior) [9]. They contribute to the regulation of social life and help individuals navigate interactions and coordinate actions, especially in unknown environments or in contexts in which belonging (or “fitting in”) is considered important [10, 17]. That norms are powerful in shaping behavior has been demonstrated in various offline contexts [8, 10, 25]. However, it has been suggested that their effects may be particularly pronounced in online contexts since (quasi) anonymity and the absence of individuating markers may lead to depersonalization, a shift from individual-focused processes to social processes [46].

Social norms may be established and spread swiftly among individuals, which may intensify unwanted, harmful, or anti-social behavior [e.g., 25], but which may also be harvested for beneficial and prosocial developments. For instance, Paluck and colleagues [36] demonstrated that training a small subset of students to speak up against conflicts improved the social climate across their entire schools. Especially trained students who were highly connected (and thus influential among their peers) changed conflict norms effectively. Similarly, in the online context, Seering and colleagues observed that influential individuals, namely moderators, shaped norms particularly well, underlining their special role in fostering desirable behavior [44]. Moreover, Bhandari and colleagues [5] showed that the visibility of moderation source information shaped the extent to which helping harassed users was seen as normative.

Returning to our systematization of forms of moderation according to facets of visibility, we expect that visibility of the offense should affect both perceptions of descriptive norms and injunctive norms, whereas visibility of counterspeech should primarily affect perceptions of injunctive norms. That is, if sexist content is visible, it should be seen as more prevalent and accepted than when it is deleted. If moderators, however, take a stance against visible sexist content, this should communicate that sexism is disapproved of.

Evidence that moderation can serve as a normative signal stems from work by Álvarez-Benjumea and Winter [1]. Using an online forum, the authors tested whether censoring of and counterspeech against hate speech would reduce the occurrence of subsequent hateful comments, assuming that the two interventions would shape descriptive and injunctive norms, respectively. The researchers found that especially moderate censoring reduced subsequent occurrence of hate speech. These results are promising since they suggest that deletion can indeed serve as a normative signal and influence future behavior; yet, it is important to note that the researchers were interested in (potential) offenders and the production of harmful content (i.e., hate speech). Still, it appears highly plausible that deletion would also constitute a normative signal for users in general. Hence, we predicted with regard to descriptive norms:

H1: *Deletion of sexist content (compared to visible sexist content) reduces the perception of sexism as prevalent.*

While Álvarez-Benjumea and Winter assumed that injunctive norms should be determined by counterspeech alone, we argue that both deletion and counterspeech may shape perceptions of injunctive norms *if* users know that deletion has occurred. In this case, the disapproval of sexist content can be inferred from deletion. If sexist content, however, remains visible, counterspeech may convey the injunctive norm that sexist content is not tolerated. Interestingly, Álvarez-Benjumea and Winter did not find that counterspeech served as an injunctive normative signal to reduce

hate speech [1]. Potentially, counterspeech is indeed less effective with regard to offenders, but it might well be effective in signaling to users in general that sexist content will be addressed and that moderators will step in for the members of their community. Consequently, we predicted an interaction effect between deletion and counterspeech on injunctive norm perceptions:

H2: *Deletion of sexist content (compared to visible sexist content) reduces the perception that sexism is accepted. If sexist content remains visible, counterspeech reduces its perceived acceptance.*

With regard to counterspeech, we moreover explored whether it reduces perceived acceptance of visible sexist content more strongly if it is public rather than private.

1.4 Feelings of Safety and Intent to Participate

To comprehensively assess what makes a specific form of content moderation effective, we argue that not only effects on offenders need to be considered [e.g., 1, 21, 22, 31, 32] but also on the larger community and in particular on individual targets or members of a targeted group, in our case, women (see [45] for a similar claim). For instance, while transparency of moderator decisions was key in the studies by Jhaver and colleagues for achieving acceptance and behavioral change in offenders, Cook and colleagues did not find that higher levels of transparency of user-driven moderation, in contrast to commercial moderation, affected general perceptions of toxicity [11]. Conversely, Miškolci and colleagues [31] found little evidence that counterspeech reduced anti-Roma sentiment on Facebook, yet it may have had bolstering effects for Roma. In other words, what is effective in reducing offenses may not necessarily be what is effective in reducing the socio-psychological toll on targeted groups. But since they are the ones who suffer psychological and societal consequences of offenses, we consider it pivotal to determine effectiveness also from their perspective.

In the present research, we define positive effects of interventions as increases in *perceived safety* and *intentions to actively participate* in an online community, which we conceptualize as the psychological and behavioral complements of the psychological harm affecting women and silencing them online. With perceived safety, we refer to the extent to which individuals see themselves as protected from harm in a given environment. Thus, perceived safety is the result of an evaluative cognitive process of environmental characteristics, and it may diverge from actual (or objective evaluations of) safety within the environment [6]. Distinct from this psychological response, intentions for participation refer to active contributions to an environment, in the context of social media especially through postings and conversing with other users. They thus constitute behavioral reactions that should reflect the extent to which individuals feel comfortable expressing themselves in a given environment.

We expected that both perceived safety and intentions for active participation are influenced by the visibility of sexism and the visibility of counterspeech in similar ways. In particular, we predicted interaction effects of both facets of visibility:

H3: *If sexist content is deleted (compared to visible), feelings of safety are higher. The adverse effect of visible sexist content is buffered by counterspeech, meaning that counterspeech (compared to no counterspeech) increases perceived safety.*

H4: *If sexist content is deleted (compared to visible), intentions for active participation are higher. The adverse effect of visible sexist content is buffered by counterspeech, meaning that counterspeech (compared to no counterspeech) increases intentions for active participation.*

Put differently, we predicted that both deletion of and counterspeech against visible sexist content would increase perceived safety and intentions for active participation. In addition, we explored

whether the effect of counterspeech is stronger if it consists of public reprimand rather than of a reference to a private message.

In principle, all individuals should benefit from the moderation of harmful content since moderation is intended to protect general rules of respect and fairness. At the same time, it seems plausible that women and men might respond somewhat differently. If sexist content is aimed at women and has the power to silence them, then they may also benefit more from its moderation, while both the sexist content and its moderation should be less impactful for men. As such, we expected that:

H3a: *The predicted effects of visibility of sexism and of counterspeech on feelings of safety are more pronounced for women than for men.*

H4a: *The predicted effects of visibility of sexism and of counterspeech on intent to actively participate are more pronounced for women than for men.*

Note that we predicted differentiated effects for men and women only with regard to feelings of safety and intent to participate, but not for norm perceptions. This is because we assumed that norm perceptions should rely more directly on the contextual information received (i.e., our manipulations) and thus should be comparable for all users, while the psychological and behavioral downstream consequences of the context configuration should also take into account the extent to which the own group (i.e., women or men) is impacted.

1.5 Moderation Evaluation

While the main focus of our research project was to investigate how the two facets of visibility determine the effectiveness of content moderation, we also considered it important to assess how the different forms of content moderation are evaluated. This approach is similar to that of Ozanne and colleagues [35][34], who investigated the effects of visibility of moderation source on both behavioral intentions and evaluations of accountability, trust, fairness, and objectivity. Considering such evaluations might be critical because perhaps what is effective may not be what social media users approve of or vice versa. For instance, while deletion may be effective in restoring feelings of safety, it may be seen as an illegitimate constraint on the freedom of speech. Such considerations resonate with the concept of *procedural justice*, which, in its essence, is concerned with the question of *how* a decision is made rather than *which* decision is made. High levels of procedural justice can increase the perceived legitimacy of an authority and the adherence to the rules it aims to uphold [52]. Recently, procedural justice has been introduced to the investigation of social media moderation and governance, with a focus on perpetrators [24, 51]. Critically, results suggest that perpetrators are less likely to violate norms in the future if they perceive rules and their application as just [24, 51]. If not only perpetrators but social media users in general care about procedural justice, we would expect that our participants judge those forms of moderation as particularly fair that provide them with information about the reasons for interventions, i.e., in particular public counterspeech. At the same time, it could be that general social media users care less about the means and more about the ends. In this case, in particular effective moderation efforts should be seen as fair.

To gain a better understanding of how participants evaluate the different forms of content moderation, we asked them to what extent they considered the moderation to be effective, fair, and trustworthy. With perceived efficacy, we refer to the potential of a form of moderation to prevent future occurrences of inappropriate content. With perceived fairness, we address the extent to which the moderation procedure and decision are seen as just and transparent. Perceived trustworthiness refers to the extent to which the moderators can be confidently tasked with the handling of inappropriate content. These parameters are loosely based on characteristics central to

the evaluation of procedural justice and legitimacy perceptions of governments and institutions [27, 52] and allow us to explore whether the two facets of visibility differentially shape justice-related evaluations.

1.6 Research Overview

To test our hypotheses, we conducted an experiment in which we asked participants to imagine joining a Facebook group. This group was supposed to be dedicated to connecting residents of their hometown, a purpose a vast number of existing Facebook groups serve and a type of group that is usually large, with both male and female members. We then presented participants with three mock posts and their associated replies from this group; two of these posts were written by female group members and were accompanied by sexist replies. We then assigned our participants randomly to one of six experimental conditions. In response to the sexist replies, each participant saw one of six forms of moderation, resulting from manipulating the two facets of visibility: With regard to visibility of sexism, we varied whether the moderator had deleted the sexist replies or not (deletion manipulation) and with regard to visibility of counterspeech, we varied whether the moderator had addressed the sexist replies publicly, privately, or not at all (counterspeech manipulation). In case of deletion as well as private counterspeech, participants were informed of the moderator's action without receiving specific information (i.e., what had been deleted or what had been said to the perpetrator).

Prior to data collection, we preregistered our hypotheses and analysis plan as well as all materials on the Open Science Framework (<https://osf.io/fduy4>). Note that we present the hypotheses here with reversed phrasing to facilitate readability; the content of the hypotheses here and in the preregistration is identical. We also uploaded our dataset and analysis script to OSF. Preregistrations are nowadays an encouraged practice in empirical research as they ensure rigorous planning of research projects and a clear differentiation between confirmatory and exploratory research questions and analyses, thereby increasing the transparency, reliability, and reproducibility of the research process and its results [28].

2 METHOD

2.1 Sample and Design

We recruited a German-speaking sample via the online platform Prolific. We planned to collect data from 840 participants, which, according to a sensitivity analysis run in G*Power [13], would allow us to detect relatively small effects ($f = .12$) with a power of .90 for interactions and main effects. In total, 841 participants completed the study. Of those, two failed the attention check (see below), and one participant requested the deletion of their data at the end of the study. Our final sample thus consisted of 839 participants, of which 412 identified as women, 413 as men, and 14 as other. Since we used gender as a factor in our main analyses and the category *other* was heavily underrepresented, we excluded it from the analyses reported here. (However, note that we also ran the central (preregistered) analyses without the factor gender on the full sample and report the results in the supplementary materials.) As such, the analyses reported here are based on a sample of 825 participants. They were on average 31.04 years old ($SD = 10.56$), spoke German as their native language or fluently, and the majority worked (57.79%) or studied (33.58%). Most participants reported using social media several times a day (55.88%) or at least daily (33.33%). We re-ran our sensitivity analysis in G*Power [13] to determine how the slightly smaller final sample size would affect the minimum size of detectable effects; changes were negligible ($f = .124$).

We used a 2 (deletion of sexist content: no vs. yes) \times 3 (counterspeech against sexist content: no vs. private vs. public) \times 2 (gender: female vs. male) between-subjects design and assigned participants randomly to conditions of the first two independent variables.

The median completion time of the study was 14 minutes, and participants received 2.25 GBP as compensation.

2.2 Procedure and Materials

After giving informed consent, participants provided socio-demographic information and responded to several scales measuring different dispositions (see the log file on OSF for further details). Next, participants were asked to imagine that they were joining a Facebook group for residents of their city, intended for any questions, requests, or offers related to the city. Participants then saw three mock Facebook posts from members of this group, each accompanied by several comments from other group members.

The first post was written by a female student searching for a room in a shared flat. The post detailed her search criteria and was accompanied by a portrait photo of a young woman, the alleged author of the post. Below, a male group member had posted an insinuating chat-up line, an example of unsolicited and inappropriate sexual advances.

The second post was written by a male group member, advertising hand-made furniture and served as a filler post.

The third and last post was written by a mother looking for a nanny for her baby since she was about to return to work. This post was accompanied by a sexist comment from a male group member, questioning why the mother does not stay at home herself and called her uncaring, a case of gender-based insult reflecting traditional gender roles.

Thus, two of the posts were accompanied by sexist comments. Depending on deletion condition, these comments were either visible or deleted; in the latter case, participants read “This message has been deleted by the moderation team”. Moreover, the sexist comments were in some cases annotated by the moderation team, depending on counterspeech conditions. In the private counterspeech condition, the annotation stated that the author had been contacted in private. In the public counterspeech condition, the author was addressed by name, explaining that the comment was degrading and not in line with the rules of the group. The author was then urged to familiarize himself with the rules and to adhere to them. Both annotations were made by a person called Alex, a name used for both women and men in German, and presented with the name affix “moderation team” so that their status was clearly communicated.

After having seen the posts, we probed participants for comprehension. We presented five statements related to the posts, and participants were asked to state whether they were true or false. If participants responded correctly, they progressed to answer dependent measures and manipulation checks. If participants responded incorrectly to any statement, the posts were presented again, and participants were urged to read them carefully. Research by Oppenheimer and colleagues [34] has shown that repeating comprehension checks for inattentive participants increases their subsequent performance, making the quality of their data comparable to that of attentive participants.

At the end of the study, participants received additional information about the purpose of the study, were thanked for their participation, and redirected to Prolific to receive their compensation.

2.3 Measures

If not stated otherwise, participants answered on rating scales ranging from 0 (*does not apply at all*) to 5 (*fully applies*). Apart from the measures reported here, we assessed several measures for further exploratory purposes (see the log file on OSF for details).




	Deletion no	Deletion yes
Counterspeech no	 <p>Klaus K. How about staying at home yourself? You're really a cruel mother!</p> <p>Like · Reply · 5h</p>  <p>Meral S. Oh how nice that it worked out with the daycare spot!</p> <p>Like · Reply · 1h</p>	 <p>Klaus K. <i>This message has been deleted by the moderation team</i></p> <p>Like · Reply · 5h</p>  <p>Meral S. Oh how nice that it worked out with the daycare spot!</p> <p>Like · Reply · 1h</p>
Counterspeech private	 <p>Klaus K. How about staying at home yourself? You're really a cruel mother!</p> <p>Like · Reply · 5h</p>  <p>Alex (Moderation Team) The author of this message has been contacted.</p> <p>Like · Reply · 5h</p>  <p>Meral S. Oh how nice that it worked out with the daycare spot!</p> <p>Like · Reply · 1h</p>	 <p>Klaus K. <i>This message has been deleted by the moderation team.</i></p> <p>Like · Reply · 5h</p>  <p>Alex (Moderation Team) The author of this message has been contacted.</p> <p>Like · Reply · 5h</p>  <p>Meral S. Oh how nice that it worked out with the daycare spot!</p> <p>Like · Reply · 1h</p>
Counterspeech public	 <p>Klaus K. How about staying at home yourself? You're really a cruel mother!</p> <p>Like · Reply · 5h</p>  <p>Alex (Moderation Team) Hello Klaus, your comment is degrading and is not in line with the rules of this group. Please make yourself familiar with them and comply with them.</p> <p>Like · Reply · 5h</p>  <p>Meral S. Oh how nice that it worked out with the daycare spot!</p> <p>Like · Reply · 1h</p>	 <p>Klaus K. <i>This message has been deleted by the moderation team.</i></p> <p>Like · Reply · 5h</p>  <p>Alex (Moderation Team) Hello Klaus, your comment is degrading and is not in line with the rules of this group. Please make yourself familiar with them and comply with them.</p> <p>Like · Reply · 5h</p>  <p>Meral S. Oh how nice that it worked out with the daycare spot!</p> <p>Like · Reply · 1h</p>

Fig. 1. Example Facebook post to illustrate the implementation of the deletion and counterspeech manipulations (English translation). On top, the post of a group member is displayed; the table below shows the sexist reply and – depending on condition – the different reactions of the moderation team.

2.3.1 Checks. Attention checks. To check whether participants were working attentively, we embedded attention checks in the ambivalent sexism inventory and the feminist identification scale. These checks instructed participants to respond by selecting “4” on the rating scale. If participants answered incorrectly, they were subsequently notified of their mistake and prompted to remain focused. We preregistered that the data of those participants who answered both attention checks incorrectly would be excluded, which was the case for only two participants.

Comprehension checks. After participants read the three mock Facebook posts, we assessed comprehension with five items for which participants had to state whether they were true or false. Three items referred to the content of posts (e.g., “In one post, someone searched for a nanny”), while two items tested whether participants had attended to information central to the manipulations (“It was mentioned that some comments had been deleted”, “One can tell whether someone is a member of the moderation team from their name”). Whether these items had to be marked as true or false depended on condition. Answering any items incorrectly prompted the repetition of the mock Facebook posts.

Manipulation checks. We used two items each to estimate whether we had successfully manipulated the visibility of sexism (e.g., “Some comments that I have read were disrespectful towards women”, $r = .97, p < .001$) and visibility of counterspeech, which referred to the public counterspeech condition in particular (e.g., “The moderation team justifies their decisions publicly”; $r = .90, p < .001$).

2.3.2 Dependent Variables. Norms. We assessed perceived group norms regarding sexism with six items. Three items measured descriptive norms (e.g., “Inappropriate comments towards women seem to be frequent”; Cronbach’s $\alpha = .75$), and three items measured injunctive norms (e.g., “Sexist comments are condoned as part of the discourse”; $\alpha = .85$).

Perceived safety and intent to participate. With regard to participants’ (imagined) membership in the group, we assessed the extent to which they felt safe in the group (e.g., “I have the impression that I can share my opinion and requests in the group without hesitation”; $\alpha = .84$) and motivated to actively use the group (e.g., “I will participate in conversations in the group”; $\alpha = .88$).

Moderation evaluation. For exploration, we assessed how participants evaluated the moderation. Three items each assessed perceived efficacy (e.g., “Moderators are efficiently tackling inappropriate comments”; $\alpha = .76$), perceived fairness (e.g., “The moderators are just in their decisions”; $\alpha = .87$), and trust (“I trust the moderation procedure”; $\alpha = .92$). Further, three items were supposed to measure perceived legitimacy (e.g., “Moderators do their job justice”) but since they did not reach acceptable reliability ($\alpha = .43$), we decided to not compute the intended scale.

2.4 Analysis Plan

We tested our hypotheses, checked the manipulations, and explored effects on moderation evaluation by means of Analysis of Variance (ANOVAs). In all analyses, we used deletion, counterspeech, and gender as independent variables. In case of interaction effects, we computed simple main effects and pairwise comparisons to which we applied Bonferroni corrections for multiple comparisons.

To explore simple associations between our measures, we first computed bivariate correlations. In addition, to gain a more comprehensive understanding of the interplay between our two facets of visibility, norm perceptions, and perceived safety and motivation to participate, we conducted mediation analyses [3, 39, 40]. The purpose of mediation analyses is to shed light on psychological mechanisms that underlie observed effects. Their general idea is that the effect of an independent variable on the dependent variable is not direct but *mediated* by a third variable (the mediator) and thus indirect. Put differently, an effect is mediated if an independent variable has an effect on the mediator, and the mediator, in turn, has an effect on the dependent variable. In our case,

perceptions of descriptive and injunctive norms were considered as mediators that should drive the effects of sexism visibility and counterspeech visibility on feelings of safety and motivation to participate. We conducted the mediation analyses using the SPSS PROCESS macro, which uses regression analyses to estimate direct and indirect effects.

3 RESULTS

3.1 Manipulation Checks

We first tested whether we had successfully manipulated the visibility of sexist content. As expected, the deletion manipulation had a large effect on the respective manipulation check, $F(1,813) = 3781.23$, $p < .001$, $\eta_p^2 = .82$, with higher perceptions of visibility in participants in visible conditions ($M = 4.61$, $SD = 0.78$) than in deletion conditions ($M = 0.70$, $SD = 1.04$). Surprisingly, we also found some further significant effects, namely a main effect of gender, $F(1,813) = 4.69$, $p = .03$, $\eta_p^2 = .01$, and interaction effects between deletion and gender, $F(1,813) = 4.85$, $p = .03$, $\eta_p^2 = .01$, counterspeech and gender, $F(2,813) = 3.51$, $p = .03$, $\eta_p^2 = .01$, and between deletion, counterspeech, and gender, $F(2,813) = 3.12$, $p = .045$, $\eta_p^2 = .01$. Critically, an inspection of these effects did not reveal clear patterns but rather small differences between single conditions of which none questioned the effectiveness of the deletion manipulation. Hence, for the sake of brevity, we report follow-up analyses on the highest-order interaction in the supplementary materials.

Next, we tested whether the manipulation of counterspeech visibility was recognized by the participants. As expected, we found a main effect of the counterspeech manipulation on the respective manipulation check, $F(2,813) = 463.48$, $p < .001$, $\eta_p^2 = .53$. Since the check addressed the public counterspeech condition, we expected the highest level of agreement from participants in that condition. Indeed, agreement in the public counterspeech condition ($M = 3.36$, $SD = 1.53$) was higher than in the no counterspeech condition ($M = 0.57$, $SD = 0.93$), $t(547) = 27.31$, $p < .001$, $d = 2.20$, and than in the private counterspeech condition ($M = 0.78$, $SD = 1.16$), $t(550) = 25.23$, $p < .001$, $d = 1.90$. Also here, we found some further significant effects, namely a main effect of deletion, $F(1,813) = 31.05$, $p < .001$, $\eta_p^2 = .04$, and significant interactions between counterspeech and deletion, $F(2,813) = 12.79$, $p < .001$, $\eta_p^2 = .03$, as well as between counterspeech and gender, $F(2,813) = 3.13$, $p = .04$, $\eta_p^2 = .01$. Please see the supplementary materials for follow-up analyses.

3.2 Norms

For descriptive norms, we found the expected main effect of deletion, $F(1,813) = 525.24$, $p < .001$, $\eta_p^2 = .39$, showing that participants perceived sexism as less prevalent if it was deleted ($M = 2.06$, $SD = 1.13$) compared to visible ($M = 3.37$, $SD = 0.69$), as predicted in H1. In addition, the main effect of gender was significant, $F(1,813) = 15.25$, $p < .001$, $\eta_p^2 = .02$, showing that women ($M = 2.92$, $SD = 1.12$) considered sexism as somewhat more prevalent than men ($M = 2.61$, $SD = 1.20$). Lastly, also the main effect of counterspeech was significant, $F(2,813) = 14.68$, $p < .001$, $\eta_p^2 = .04$, which was further qualified by an interaction with deletion, $F(2,813) = 30.75$, $p < .001$, $\eta_p^2 = .07$ (see Fig. 2 left panel).

Follow-up simple main effects computed per deletion condition showed that descriptive norm perceptions did not differ depending on counterspeech in the no deletion condition, $F(2,813) = 1.82$, $p = .16$, $\eta_p^2 = .004$. In the deletion condition, instead, the effect of counterspeech was significant, $F(2,813) = 43.48$, $p < .001$, $\eta_p^2 = .10$. Post-hoc comparison showed that sexism was seen as more prevalent if sexism was deleted yet publicly reprimanded both compared to no counterspeech, $t(275) = 5.22$, $p < .001$, $d = 0.86$, and private counterspeech, $t(271) = 4.63$, $p < .001$, $d = 0.82$. All other interactions were non-significant, $ps > .06$.

With regard to perceived injunctive norms, we found main effects of both deletion, $F(1,813) = 480.08$, $p < .001$, $\eta_p^2 = .37$, and counterspeech, $F(2,813) = 231.90$, $p < .001$, $\eta_p^2 = .36$. As predicted in H2, also their interaction was significant, $F(2,813) = 181.18$, $p < .001$, $\eta_p^2 = .31$. In addition, also the counterspeech \times gender interaction was significant, $F(2,813) = 3.32$, $p = .04$, $\eta_p^2 = .01$ (see Fig. 2 right panel). An inspection of pairwise comparisons revealed that this is only due to women ($M = 2.43$, $SD = 1.65$) seeing sexism as somewhat more accepted than men ($M = 2.12$, $SD = 1.62$), if there is no counterspeech, $t(271) = 2.17$, $p = .03$, $d = .19$. Other than that, the pattern of results regarding injunctive norms are highly similar for men and women.

Following up on the predicted deletion \times counterspeech interaction, simple main effects showed that counterspeech affected injunctive norm perceptions if sexist comments remained visible, $F(2,813) = 405.21$, $p < .001$, $\eta_p^2 = .50$, but not if they were deleted, $F(2,813) = 2.53$, $p = .08$, $\eta_p^2 = .01$. If visible and not addressed, sexism was perceived as considerably more accepted, compared to if it was addressed privately, $t(271) = 15.03$, $p < .001$, $d = 2.25$, or publicly, $t(270) = 20.19$, $p < .001$, $d = 3.34$. Perceptions of acceptance also differed between the two forms of counterspeech and were lower after public counterspeech, compared to private counterspeech, $t(277) = -5.17$, $p < .001$, $d = -0.56$. Thus, in line with H2, sexism was seen as less accepted if it remained visible but was addressed, especially if addressed publicly.

Taken together, deletion and counterspeech shaped descriptive and injunctive norms largely in line with our predictions: Deletion of sexist comments led to the perception that sexism was less prevalent; this effect was dampened if deletion was coupled with a public reprimand. Conversely, deletion generally communicated that sexist comments were not tolerated; if sexist comments remained visible, also counterspeech evoked a sense that sexism is not tolerated, though especially public counterspeech approximated the effectiveness of deletion in reducing perceived acceptance.

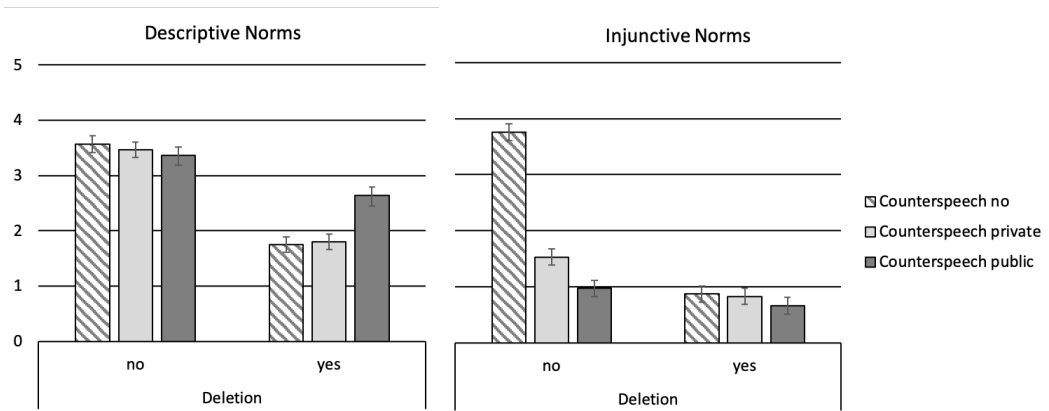


Fig. 2. Effects of deletion and counterspeech on perceived descriptive and injunctive norms. Deletion (especially if coupled with private or no counterspeech) reduced perceived prevalence of sexist content. Deletion also reduced perceived acceptance of sexist content; visible sexist content was perceived as less accepted after private or public counterspeech. Error bars represent 95% confidence intervals.

3.3 Perceived Safety and Intent to Participate

Next, we tested whether deletion and counterspeech would increase feelings of safety, especially for women. We found main effects of deletion, $F(1,813) = 90.52$, $p < .001$, $\eta_p^2 = .10$, counterspeech, $F(2,813) = 13.72$, $p < .001$, $\eta_p^2 = .03$, and gender, $F(1,813) = 6.16$, $p = .01$, $\eta_p^2 = .01$. Importantly, as

predicted in H3, the deletion \times counterspeech interaction was significant, $F(2,813) = 12.38$, $p < .001$, $\eta_p^2 = .03$; yet – contrary to H3a – this was not further qualified by gender, $F(2,813) = 0.45$, $p = .64$, $\eta_p^2 = .001$. Gender did, however, interact with deletion, $F(1,813) = 12.12$, $p < .001$, $\eta_p^2 = .02$.

We followed up on the interaction effects by computing simple main effects of counterspeech and gender separately for each deletion condition.

Counterspeech affected feelings of safety in the no deletion condition, $F(2,813) = 25.72$, $p < .001$, $\eta_p^2 = .06$, but not in the deletion condition, $F(2,813) = 0.08$, $p = .92$, $\eta_p^2 < .001$ (see Fig. 3, left panel). Pairwise comparisons in the no deletion condition showed private, $t(271) = 4.51$, $p < .001$, $d = 0.53$, and public, $t(270) = 7.10$, $p < .001$, $d = 0.83$, counterspeech could increase feelings of safety, compared to no counterspeech. Feelings of safety also differed between the two types of counterspeech and were slightly higher if sexist comments were addressed publicly, $t(277) = 2.63$, $p = .03$, $d = 0.31$. Put differently, if sexism remained visible, participants felt safest if moderators addressed it publicly. Still, public counterspeech in combination with deletion achieved higher feelings of safety than public counterspeech with sexism remaining visible, $t(274) = 2.21$, $p = .03$, $d = 0.26$. These effects are illustrated in Fig. 3.

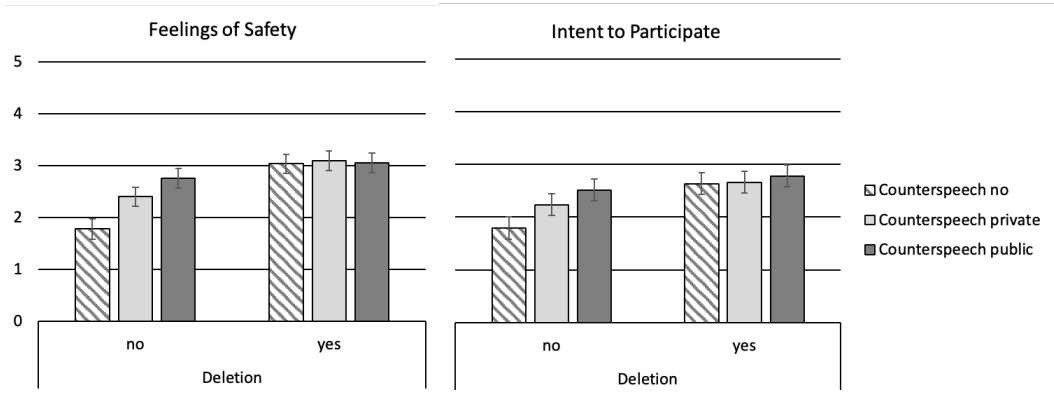


Fig. 3. Effects of deletion and counterspeech on feelings of safety and motivation to participate. Deletion increased feelings of safety and intent to participate; if sexist content was visible, especially public counterspeech increased feelings of safety and intent to participate. Error bars represent 95% confidence intervals.

With regard to gender, only the simple main effect in the no deletion condition was significant, $F(1,813) = 17.75$, $p < .001$, $\eta_p^2 = .02$ (deletion condition, $F(1,813) = 0.50$, $p = .48$, $\eta_p^2 = .001$). If sexist comments remained visible, women felt less safe ($M = 2.08$, $SD = 1.23$) than men ($M = 2.55$, $SD = 1.23$), in line with the idea that women are silenced by sexist content on social media.

The pattern of results for intent to participate in the online debate was highly similar. We again found main effects of deletion, $F(1,813) = 35.62$, $p < .001$, $\eta_p^2 = .04$, and counterspeech, $F(2,813) = 8.55$, $p < .001$, $\eta_p^2 = .02$ (though not of gender, $F(1,813) = 0.99$, $p = .34$, $\eta_p^2 = .001$). Critically, supporting H4, these main effects were qualified by a significant deletion \times counterspeech interaction, $F(2,813) = 3.99$, $p = .02$, $\eta_p^2 = .01$; yet again – against H4a – this was not further qualified by gender. Instead, the deletion \times gender interaction was again significant, $F(1,813) = 8.72$, $p = .003$, $\eta_p^2 = .01$.

Computing simple main effects per deletion condition showed that the effect of counterspeech was only significant in the no deletion condition, $F(2,813) = 11.85$, $p < .001$, $\eta_p^2 = .03$, but not in the deletion condition, $F(2,813) = 0.56$, $p = .57$, $\eta_p^2 = .001$ (Fig. 3 right panel). Pairwise comparisons of counterspeech conditions if sexist comments remained visible revealed that the intent to participate

in the group was higher if sexist comments were addressed privately, $t(271) = 2.96$, $p = .01$, $d = 0.35$, or publicly, $t(270) = 4.83$, $p < .001$, $d = 0.55$, compared to no counterspeech. Intent did not differ between private and public counterspeech, $t(277) = 1.89$, $p = .18$, $d = 0.23$. Comparing the effect of public counterspeech between deletion conditions showed that it fared comparably well in both conditions, $t(274) = 1.81$, $p = .07$, $d = 0.22$. Private counterspeech, instead, motivated participation more if sexist comments were deleted rather than visible, $t(274) = 2.86$, $p = .004$, $d = 0.34$.

The deletion \times gender interaction revealed an interesting pattern of results: Motivation to participate in the group did not differ if sexism remained visible, $F(1,813) = 1.99$, $p = .16$, $\eta_p^2 = .002$, but when it was deleted, men's motivation to participate was lower ($M = 2.52$, $SD = 1.17$) than women's ($M = 2.86$, $SD = 1.20$), $F(1,813) = 7.66$, $p = .01$, $\eta_p^2 = .01$.

Thus, taken together, deletion – irrespective of counterspeech – increased feelings of safety and the motivation to actively use the Facebook group. If sexism remained visible, then public reprimand of sexist comments, in particular, showed to be effective.

3.3.1 Mediation Analyses. Do counterspeech and deletion exert their effects on feelings of safety and intent to actively use the Facebook group by altering social norms? To test this, we first explored bivariate correlations (Table 1). In line with our theoretical reasoning, we found that both feelings of safety and intent to participate were lower the more sexist behavior was seen as prevalent and as accepted.

	1	2	3	4	5	6
1 Descriptive Norms	–					
2 Injunctive Norms	.40**	–				
3 Perceived Safety	-.38**	-.40**	–			
4 Intent to Participate	-.20**	-.25**	.75**	–		
5 Efficacy	-.43**	-.68**	.46**	.30**	–	
6 Fairness	.08*	-.49**	.30**	.23**	.49**	–
7 Trustworthiness	-.15**	-.65**	.43**	.32**	.71**	.80**

* $p < .05$, ** $p < .01$

Table 1. Bivariate correlations between central and exploratory measures.

Next, bringing all components together, we computed two mediation models using the SPSS PROCESS macro. In each model, we used counterspeech (dummy-coded, D1: no counterspeech vs. private counterspeech; D2: no counterspeech vs. public counterspeech, with no counterspeech being coded 0) and deletion as predictors, and descriptive norms and injunctive norms as parallel mediators. Note that since we did not find evidence for the expected differentiation of the deletion \times counterspeech interactions on safety and intent depending on gender, we did not consider gender in these analyses. In the first model, we input feelings of safety as the outcome and in the second intent to participate. We used Process Model 8 in which interaction effects of the predictors on both the mediators and the outcome are considered. In case of significant interactions, the counterspeech effects were estimated separately for the deletion and no deletion condition. (Note that in Process, deletion is thus, in fact, defined as a moderator. For simplicity, we stick to the term predictor.) The mediation model for perceived safety is depicted in Fig. 4 and the model for intent to participate in Fig. 5.

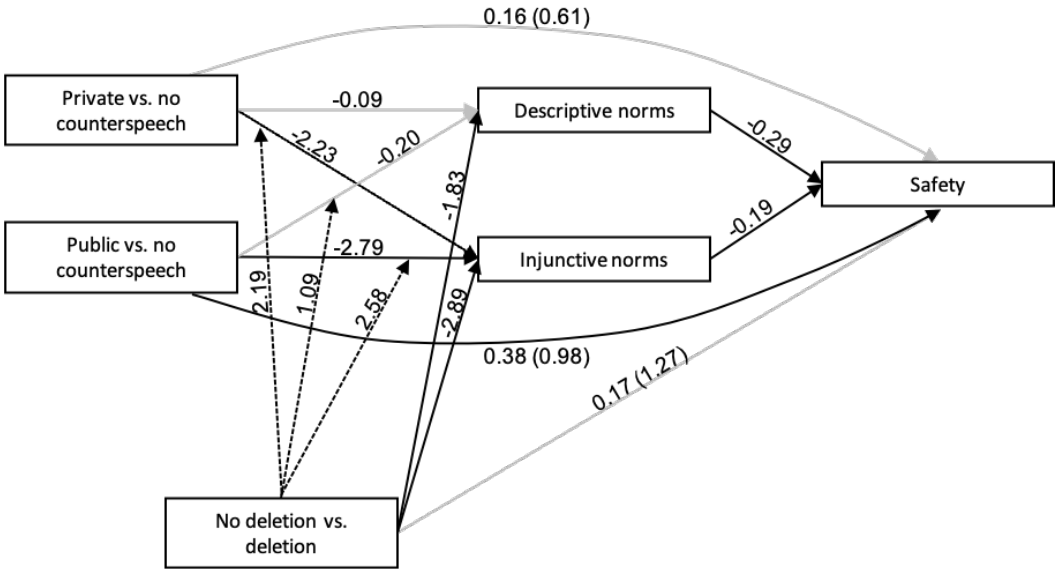


Fig. 4. Mediation Analysis 1 – predicting perceived safety from deletion and counterspeech, mediated by descriptive and injunctive norms. Solid black lines depict significant paths, grey lines non-significant paths. Dotted lines depict significant interaction effects. Estimates are non-standardized, with total effects in brackets.

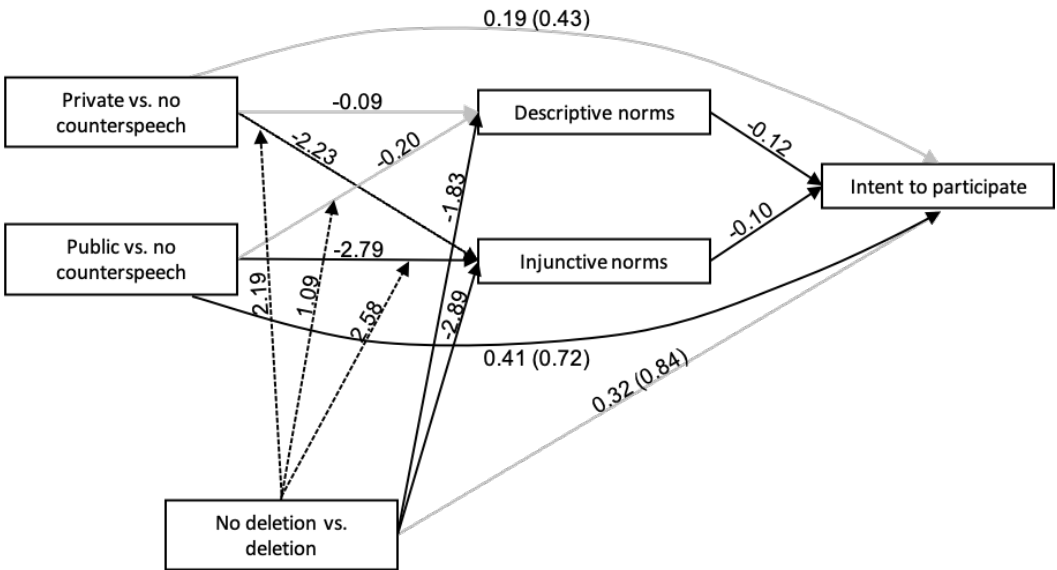


Fig. 5. Mediation Analysis 2 – predicting intent to participate from deletion and counterspeech, mediated by descriptive and injunctive norms. Solid black lines depict significant paths, grey lines non-significant paths. Dotted lines depict significant interaction effects. Estimates are non-standardized, with total effects in brackets.

The mediation analyses provided a rather consistent pattern of indirect effects for both feelings of safety and intent to participate (Table 2). Deletion affected feelings of safety through changes in descriptive norms and injunctive norms. The effect of deletion on intent to participate, instead, was only mediated through descriptive norms (see rows 1 and 2 of Table 2). The effects of private as well as public counterspeech on perceived safety were mediated through changes in injunctive norms and – interestingly – in the case of public counterspeech also through descriptive norms. With regard to intent to participate, the effect of private counterspeech was mediated through injunctive norms and of public counterspeech through descriptive norms (rows 3 to 6 of Table 2).

As depicted in Fig. 4 and Fig. 5, the effects of counterspeech on descriptive and injunctive norms were qualified by deletion. Thus, we also looked at the conditional indirect effects of counterspeech separately for each deletion condition. This revealed that in particular the effect of counterspeech against visible sexist comments on safety was mediated through norms (rows 7 to 10 of Table 2).

In sum, this suggests that social norms seem to be an important mechanism underlying the effects of deletion and counterspeech on feelings of safety and intent to participate. Both may affect feelings of safety and intent to participate by lowering the perceived prevalence and acceptance of sexist content. Effects of counterspeech were in particular driven by changes in norm perceptions if sexist content remained visible.

	Safety			Intent to Participate		
	Effect	SE	95% CI	Effect	SE	95% CI
Deletion → descriptive norms	.54	.09	.37, .73	.23	.09	.03, .42
Deletion → injunctive norms	.56	.13	.28, .85	.29	.15	-.003, .59
Private → descriptive norms	.03	.03	-.02, .08	.01	.01	-.01, .05
Private → injunctive norms	.43	.10	.21, .66	.23	.12	.002, .46
Public → descriptive norms	.06	.03	.01, .12	.03	.02	.003, .07
Public → injunctive norms	.54	.13	.27, .88	.28	.14	-.001, .58
Conditional indirect effects:						
Deletion no						
Private → descriptive norms	.03	.03	-.02, .08	.01	.01	-.01, .04
Private → injunctive norms	.43	.11	.22, .65	.23	.12	-.005, .46
Public → descriptive norms	.06	.03	.01, .12	.03	.02	.001, .06
Public → injunctive norms	.54	.14	.28, .81	.28	.15	-.01, .57
Conditional indirect effects:						
Deletion yes						
Private → descriptive norms	-.02	.04	-.10, .06	-.01	.02	-.05, .03
Private → injunctive norms	.01	.02	-.03, .04	.004	.01	-.02, .03
Public → descriptive norms	-.26	.06	-.38, -.16	-.11	.05	-.21, -.02
Public → injunctive norms	.04	.02	.01, .09	.02	.02	-.001, .06

Table 2. Indirect effects (bootstrapped, 5000 samples) of deletion and counterspeech on safety and motivation via descriptive and injunctive norms. Significant paths are printed in bold.

3.4 Moderation Evaluation

Lastly, we explored the extent to which participants evaluated the different forms of moderation as effective, trustworthy, and fair.

3.4.1 Efficacy. We found significant main effects of deletion, $F(1,813) = 405.92, p < .001, \eta_p^2 = .33$, counterspeech, $F(2,813) = 73.46, p < .001, \eta_p^2 = .15$, and gender, $F(1,813) = 5.60, p = .02, \eta_p^2 = .01$, on efficacy. The significant effect of gender showed that men ($M = 2.97, SD = 1.30$) found all procedures somewhat more effective than women ($M = 2.74, SD = 1.28$).

The main effects of deletion and counterspeech were further qualified by a significant interaction, $F(2,813) = 54.93, p < .001, \eta_p^2 = .12$. This is visualized in Fig. 6, upper panel. If sexist comments were deleted, this was seen as rather effective, irrespective of additional counterspeech, $F(2,813) = 1.54, p = .22, \eta_p^2 = .004$. If sexism remained visible, both types of counterspeech were seen as similarly effective, $t(277) = 1.41, p = .48, d = 0.15$, and clearly more effective than no counterspeech; for private counterspeech, $t(271) = 13.04, p < .001, d = 1.40$, for public counterspeech, $t(270) = 14.44, p < .001, d = 1.59$. Still, both forms of counterspeech were seen as more effective in combination with deletion than if sexist comments remained visible, for private counterspeech, $t(274) = 8.70, p < .001, d = 1.01$; for public counterspeech, $t(274) = 6.10, p < .001, d = 0.74$.

Taken together, comparing the different forms of moderation, any form of deletion – with or without counterspeech – was seen as rather effective. Counterspeech against visible sexist content was also seen as somewhat effective, yet not to the same extent as deletion.

3.4.2 Trustworthiness. Regarding trustworthiness, we found significant main effects of deletion, $F(1,813) = 91.40, p < .001, \eta_p^2 = .10$, counterspeech, $F(2,813) = 156.25, p < .001, \eta_p^2 = .28$, and gender, $F(1,813) = 5.38, p = .02, \eta_p^2 = .01$. Also the 2-way interactions between deletion and counterspeech, $F(2,813) = 87.34, p < .001, \eta_p^2 = .18$, and deletion and gender, $F(1,813) = 18.63, p < .001, \eta_p^2 = .02$, were significant. Lastly, also the deletion \times counterspeech \times gender interaction was significant, $F(2,813) = 3.55, p = .03, \eta_p^2 = .01$. The effect is visualized in Fig. 6, middle panel.

To follow up on the 3-way interaction, we looked at the effects of deletion and counterspeech on perceived trustworthiness for men and women separately.

For women, trust did not differ depending on counterspeech if sexist comments were deleted, $F(2,813) = 0.55, p = .55, \eta_p^2 = .001$, but it differed if sexist comments remained visible, $F(2,813) = 140.80, p < .001, \eta_p^2 = .26$. Unsurprisingly, this effect was mainly driven by the no counterspeech condition in which the work of moderators was seen as considerably less trustworthy, compared to private counterspeech, $t(143) = 13.59, p < .001, d = 2.14$, and public counterspeech, $t(138) = 15.48, p < .001, d = 2.94$. Whether (visible) sexist comments were publicly or privately annotated did not affect perceptions of trustworthiness, $t(145) = 2.08, p = 1.00, d = 0.35$. Moreover, private counterspeech led to more trust when it was coupled with deletion, $t(138) = 2.54, p = .01, d = 0.43$. Public counterspeech, instead, was equally effective with sexist comments being visible or deleted, $t(133) = 0.20, p = .85, d = 0.04$.

For men, instead, trust differed depending on counterspeech in both the deletion condition, $F(2,813) = 8.03, p < .001, \eta_p^2 = .02$, and no deletion condition, $F(2,813) = 97.85, p < .001, \eta_p^2 = .19$. Also here, if sexist comments remained visible and were not annotated, trust was lower than with private, $t(126) = 10.63, p < .001, d = 1.65$, or public counterspeech, $t(130) = 13.26, p < .001, d = 2.19$. Public counterspeech also elicited somewhat more trust than private counterspeech, $t(130) = 2.45, p = .045, d = 0.43$. A similar pattern emerged if sexist comments were deleted. Men's trust was higher if the deletion was accompanied by a public counterspeech compared to no counterspeech, $t(143) = 3.99, p < .001, d = 0.62$. Trust in the private counterspeech condition differed neither from trust in the public counterspeech condition, $t(143) = 1.74, p = .24, d = 0.30$, nor in the no counterspeech condition, $t(142) = 2.25, p = .08, d = 0.33$. Lastly, if sexist comments were publicly annotated, trust

was equally high, irrespective of whether the comments were deleted or not, $t(139) = 1.34$, $p = .18$, $d = 0.24$. The same was true for private counterspeech, $t(134) = 0.53$, $p = .60$, $d = 0.09$.

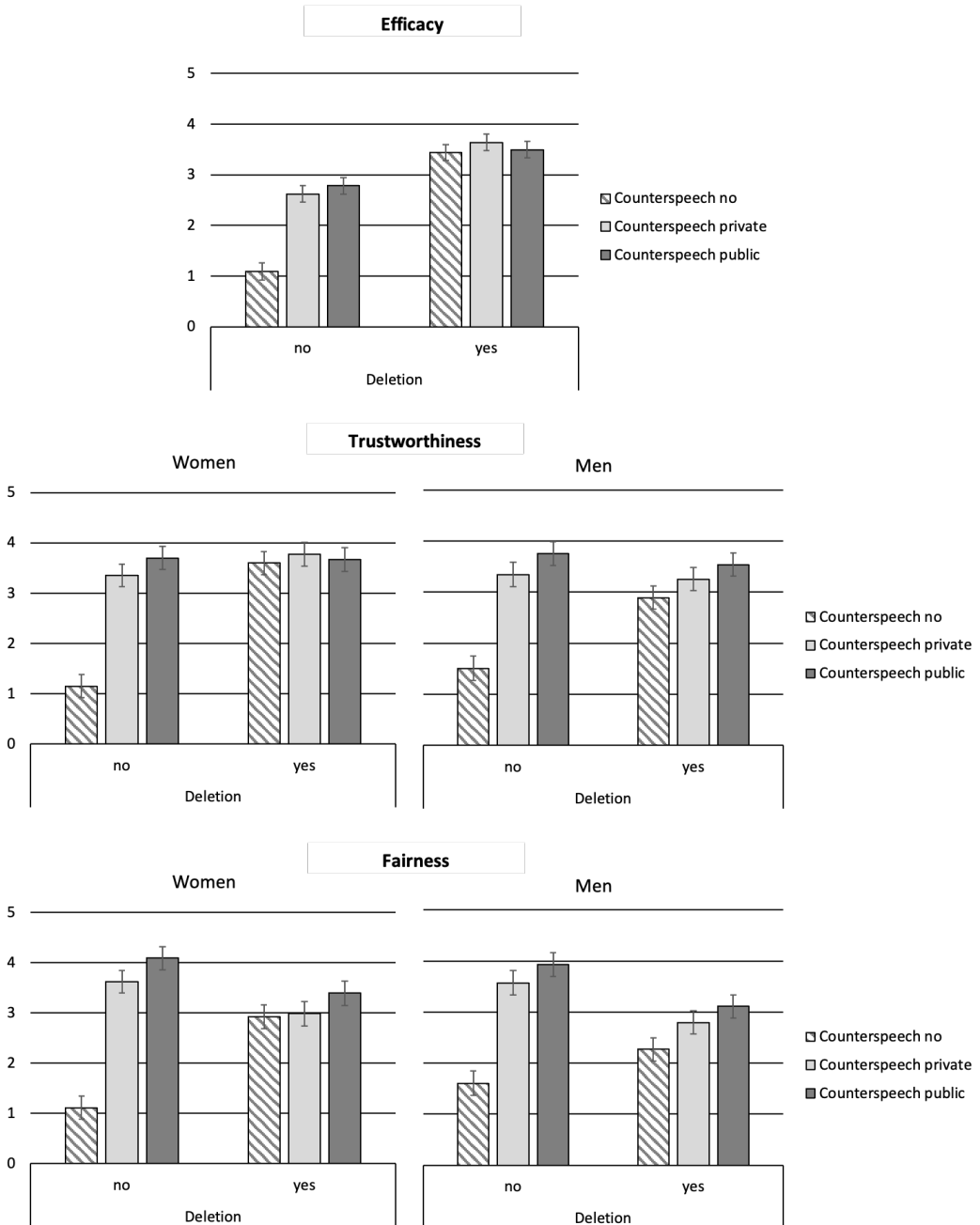


Fig. 6. Effects of deletion and counterspeech on moderation efficacy, trustworthiness, and fairness. Effects for trustworthiness and fairness are presented separately for female and male participants to account for respective significant interactions. Error bars represent 95% confidence intervals.

In sum, despite slight variations in the patterns of results for women and men, all participants, on average, trusted all forms of moderation to a considerable extent and substantially more than when the moderator remained inactive in response to sexist content.

3.4.3 Fairness. For perceptions of fairness, we found main effects of counterspeech, $F(2,813) = 211.16$, $p < .001$, $\eta_p^2 = .34$, and gender, $F(1,813) = 5.23$, $p = .04$, $\eta_p^2 = .01$, significant 2-way interactions between deletion and counterspeech, $F(2,813) = 91.02$, $p < .001$, $\eta_p^2 = .18$, and deletion and gender, $F(1,813) = 11.68$, $p < .001$, $\eta_p^2 = .01$, as well as a significant 3-way interaction deletion \times counterspeech \times gender, $F(2,813) = 5.97$, $p = .003$, $\eta_p^2 = .01$. The effect is visualized in Fig. 6, lower panel.

Again, we looked at the effects of deletion and counterspeech separately for women and men. For women, perceived fairness differed with counterspeech both if sexist comments remained visible, $F(2,813) = 184.96$, $p < .001$, $\eta_p^2 = .31$, or were deleted, $F(2,813) = 4.45$, $p = .01$, $\eta_p^2 = .01$. If sexist comments remained visible, fairness perceptions were lowest if they were not annotated, compared to private counterspeech, $t(143) = 15.29$, $p < .001$, $d = 2.42$; and compared to public counterspeech, $t(138) = 17.91$, $p < .001$, $d = 3.37$. Moreover, public counterspeech was seen as fairer than private counterspeech, $t(145) = 2.88$, $p = .01$, $d = 0.54$. If sexist comments were deleted, public counterspeech was seen as fairer than no counterspeech, $t(130) = 2.77$, $p = .02$, $d = 0.48$. Fairness perceptions after private counterspeech neither differed from no counterspeech, $t(130) = 0.37$, $p > .99$, $d = 0.06$, nor from public counterspeech, $t(126) = 2.36$, $p = .06$, $d = 0.38$. Interestingly, comparing the effects of counterspeech between conditions of deletion revealed that both private counterspeech, $t(138) = 3.82$, $p < .001$, $d = 0.61$, and public counterspeech, $t(133) = 4.08$, $p < .001$, $d = 0.76$, were perceived as fairer in the no deletion condition.

For men, the overall pattern of results was similar, yet, more pronounced: Differences between conditions of counterspeech were significant in both the no deletion condition, $F(2,813) = 105.30$, $p < .001$, $\eta_p^2 = .21$, and the deletion condition, $F(2,813) = 13.50$, $p < .001$, $\eta_p^2 = .03$. If sexist comments remained visible and were not annotated, fairness perceptions were markedly lower than after private counterspeech, $t(126) = -11.32$, $p < .001$, $d = -1.83$, and public counterspeech, $t(130) = 13.61$, $p < .001$, $d = 2.13$. Perceived fairness did not differ between public and private counterspeech, $t(126) = 2.09$, $p = .11$, $d = 0.39$. Also, if sexist comments had been deleted, then no counterspeech was seen as less fair than private counterspeech, $t(142) = 3.22$, $p = .004$, $d = 0.53$, and public counterspeech, $t(143) = 5.13$, $p < .001$, $d = 0.87$. Again, perceived fairness did not differ between public and private counterspeech, $t(143) = 1.91$, $p = .17$, $d = 0.33$. Critically, just as for women, both private counterspeech, $t(134) = 3.82$, $p < .001$, $d = 0.80$, and public counterspeech, $t(139) = 4.08$, $p < .001$, $d = 0.93$, were perceived as fairer if sexist comments remained visible.

Thus, strikingly, fairness evaluations of both private and public counterspeech were reduced if they were coupled with deletion (compared to when sexist content remained visible). This pattern could be observed for both men and women but was particularly pronounced for men.

4 DISCUSSION

Sexist content is widespread on social media and has been shown to have detrimental psychological and social effects. In the here presented research, we investigated the effectiveness and acceptance of several forms of moderation that vary in the extent to which sexist content remains visible and in the extent to which counterspeech is visible. By considering both effectiveness and acceptance, we are able to identify alignments and potential tradeoffs that may shape the success of a form of moderation in the long run. Critically, we determined effectiveness and acceptance from the perspective of users in general and in particular from the perspective of women as members of the

targeted group, thereby putting the wider community and targets in the foreground rather than perpetrators.

4.1 Effectiveness of Different Forms of Moderation

In an online experiment, we presented participants with a (mock) Facebook group in which they saw sexist replies to postings of female group members. Depending on experimental condition, a moderator had reacted to these sexist replies in different ways. To manipulate the visibility of sexism, we varied whether the moderator had deleted the sexist replies or not; to manipulate visibility of counterspeech, we varied whether the moderator had engaged in public, private, or no counterspeech. As predicted in H1 and H2, we found that deletion of sexist content (compared to visible sexist content) reduced the perception that sexism is prevalent (i.e., descriptively normative) and accepted (i.e., injunctively normative) in the group. If sexist content remained visible, its perceived acceptance was reduced by counterspeech, especially if it occurred publicly.

Similarly, we found that the deletion of sexist content increased feelings of safety and intent to actively participate in the group. If sexist content remained visible, in particular public counterspeech could effectively increase feelings of safety and intent to participate. These results support H3 and H4. Thus, deletion served as both a descriptive and an injunctive normative signal and increased feelings of safety and intent to participate, irrespective of whether it was accompanied by counterspeech or not. In other words, whether the moderator additionally reprimanded the perpetrator privately or publicly was of little importance *if* sexist content had been deleted. Conversely, if sexist content remained visible, counterspeech – and in particular public counterspeech – served as an injunctive normative signal and could increase feelings of safety and intent to participate.

These results tie in well with the findings by Im and colleagues [20] on women's preferences for moderation. The authors had identified the removal of insulting and disrespecting comments as women's preferred form of moderation, followed by labeling these comments as such. Beyond preference, our findings show that these forms of moderation can also be effective.

Our exploratory mediation analyses allowed us to gain a more comprehensive understanding of the psychological processes at play. We found that the effects of deletion on feelings of safety and intent to participate were mediated through perceived descriptive (and injunctive norms, in the case of safety). If sexist content remained visible, the effect of counterspeech on feelings of safety and intent to participate was primarily mediated through injunctive norms. This suggests that deletion and counterspeech can effectively change what is perceived as normative, which, in turn, shapes psychological and behavioral reactions. These findings are in line with a growing body of literature showing that norms as set by moderators may be a critical driver of attitudes and behavior in social media settings [1, 44, 45] and adds to existing work the perspective of the wider community and targeted groups.

While we had predicted that the interplay of deletion and counterspeech should produce a more pronounced pattern of results for women than for men, this was not the case. However, we did find that women felt less safe than men if sexist comments remained visible and, conversely, that women were more motivated to actively participate in the group than men if sexism was deleted. This suggests that women might indeed have been impacted more strongly than men by the presence or absence of deletion (but not of counterspeech). The fact that also men's feelings of safety and intent to participate were considerably shaped by deletion and counterspeech suggests that the moderator's actions might have served as examples that show how they deal with misconduct in general (not only against women), thus shaping a general sense of safety.

Our findings suggest that even a few instances of moderation of sexist content can have a considerable impact on users' perceived safety and their intent to actively participate in an online environment. While all forms of moderation fared better than no moderation at all, deletion and

public counterspeech against visible sexist content were especially effective. This underlines how impactful actions of moderators can be for the functioning of online communities, not only in terms of regulating wrongful conduct but also in terms of creating an environment in which users feel safe to speak up.

4.2 Evaluation of Different Forms of Moderation

Our exploratory analyses of moderation evaluation with regard to efficacy, trustworthiness, and fairness revealed a differentiated and highly interesting picture. In general, participants considered moderation most effective if it involved the deletion of sexist content. Its trustworthiness, instead, was rather high as long as *any* action was taken, with some small differences between different forms of moderation. Fairness, on the other hand, was highest for counterspeech against visible rather than deleted sexist content, especially if counterspeech occurred publicly.

These results, first and foremost, shed an interesting light on deletion: While participants considered counterspeech against deleted sexist content effective in reducing future occurrences – and it *was* effective in increasing safety and intent to participate – they appeared to have questioned its fairness. Why is this the case? Perhaps for moderation to be considered fair, users do not only require transparency with regard to the moderator’s motive (which is given in conditions in which deletion is coupled with counterspeech) but also with regard to the offense. Only if users themselves can judge the offense, then they can determine whether actions taken by the moderators are justified and fair or whether – in their view – a perpetrator has been treated wrongfully. Such an interpretation would offer an interesting extension to the findings by Jhaver and colleagues [22]. While they showed that transparency of the moderator’s motives matters for perpetrators, our results seem to suggest that also members of the wider community value transparency and not only with regard to the moderator’s motives but also with regard to the offense. Further research is needed to substantiate this interpretation and could, for instance, explore whether giving users the option to make deleted content visible again for their review could boost fairness perceptions. Such an option could consist of an additional mouse click, similar to the Twitter option¹ to make sensitive media visible. Still, the observation that counterspeech against deleted sexist content is considered less fair suggests that such moderation could backfire.

4.3 Practical Implications

For social media platforms that intend to improve their moderation practices to create a safer space for their users, our experiment yields valuable insights. First, deleting sexist comments was effective in increasing feelings of safety and intentions to participate. This was true irrespective of whether deletion was accompanied by counterspeech or not, suggesting that a “cheap” form of moderation in which sexist content is merely deleted is just as effective as more complex forms of moderation. At the same time, it is critical to note that deletion was seen as less fair than counterspeech against visible sexist content, suggesting that there may be a possible tradeoff between effectiveness and fairness. If, in the long run, this tradeoff tilts towards valuing fairness, this could reduce the effectiveness of deletion.

We would also like to add a cautionary note on the implementation of deletion. In our study, deletion did not happen covertly, but users were informed whenever it had taken place. This provided them with both the information that some users share sexist content and that there are moderators who delete it, which is probably why deletion did not only shape descriptive but also

¹See: Twitter’s sensitive media policy at <https://help.twitter.com/en/rules-and-policies/media-policy>. Retrieved on January 12, 2023.

injunctive norms. We argue that informing users whenever content has been deleted is important to avoid the illusion that sexism does not exist in an environment.

If sexist content remained visible, in particular public counterspeech showed to be effective in increasing feelings of safety and intent to participate. The fact that public counterspeech fared better than private counterspeech may again be a matter of transparency; only in the public condition participants learned why and how perpetrators were sanctioned. In other words, it might not only matter *that* sexist content was addressed but also *how* it was addressed. Moreover, public counterspeech was regarded as particularly fair.

Social media platforms and their moderators need to weigh a range of aspects when deciding how to deal with sexist content, yet from the perspective of the wider user community, our research suggests that especially public counterspeech may be both effective in creating a safe space and regarded as fair. It also stands in less conflict with the right to free speech, compared to deletion, and could thus be applied to cases of sexist behavior that do not violate legal norms but still cause considerable psychological harm.

Our finding that social norms are a driving factor for creating a safe space also provides insights for dealing with sexist content beyond moderation. While moderators may be particularly powerful in shaping norms [44], also ordinary users can take a stance against sexist content [42], which would presumably establish the injunctive norm that sexism is not tolerated. Indirect evidence for this stems from work by Mathew and colleagues [30], who showed that user-generated counterspeech on YouTube received more likes than neutral comments. As such, social media platforms may also seek out ways in which their community can be motivated to engage in counterspeech to utilize injunctive norms as a path to a safer environment.

Our work was situated in the context of Facebook groups in which human moderators monitor content and intervene if deemed necessary, yet it also needs to be considered in light of the increasing automation of moderation [e.g., 12, 18]. Presupposing accurate detection of sexist content, its subsequent deletion can be swiftly implemented to create a safer environment. Critically, while our results suggest that deletion may lack fairness, Gonçalves and colleagues [16] found that deletion without further explanation was perceived as more transparent if executed automatically rather than by a person. Thus, automated deletion could perhaps overcome the observed lack of fairness and constitute a viable and efficient form of moderation. Apart from deletion, researchers have also begun to develop sophisticated systems to create counterspeech [12, 41, 56]. While this endeavor may have been mainly motivated by automating a form of moderation that ensures free speech, our results stress that counterspeech is also effective in creating a safer environment and perceived as fair.

Taken together, our research provides novel insights into the moderation of sexist content from the perspective of the broad user community and in particular from women as members of the targeted group. Further, our findings can inform the governance and development of social media platforms. Our insights on the effectiveness and evaluation of deletion and counterspeech may be taken into account in the training of human moderators and in the development of automated moderation.

4.4 Limitations and Future Research

While our study provides several meaningful insights, it is not without limitations. First, we only presented our participants with a one-time snapshot of a (mock) Facebook group. It is plausible that users are less reflective and engage in fewer evaluative processes when scrolling through their social media feeds than we prompted our participants to do in the study. Still, we regarded a highly controlled study as a necessary first step of investigation that can inform future studies with higher ecological validity. Relatedly, with regard to the active use of the group, we only assessed

participants' self-reported intentions. Yet, intentions do not always translate into behavior [4], and future research should aim to assess actual user behavior. Lastly, it is plausible that potential adverse reactions to some forms of moderation only occur over time. For instance, while participants judged deletion to be rather effective in reducing the occurrence of sexist content, if such a reduction does not manifest over time, efficacy judgments might go down.

Apart from increasing ecological validity, future research should also broaden the scope of investigation. The focus of our research rested on the perspective of general users and members of the targeted group in particular. To fully stake out the effectiveness of the tested forms of moderation, future research should refine and broaden this perspective by considering effects on direct targets as well as perpetrators. For instance, our results showed that public counterspeech seems to work well for users, yet it might be perceived as public shaming by perpetrators, which could cause them to strike back or polarize more. Lastly, we focused on the moderation of sexist content since this is a particularly frequent and impactful form of misconduct in online settings. Yet, future research should test whether the effects of deletion and counterspeech replicate in response to other forms of misconduct or are, to some extent, context-specific.

4.5 Conclusion

Our research provides important insights into the effectiveness and acceptance of different forms of moderation in reaction to sexist content on social media from the perspective of the user community. We could demonstrate that both female and male users felt safer and more inclined to contribute to their community if sexist content was deleted or reprimanded, especially if it was reprimanded publicly, suggesting that the *invisibility* of sexist content or the visibility of counterspeech matter for creating a safe online environment. These effects were driven by changes in social norm perceptions. At the same time, deletion was considered less fair than a reprimand of visible sexist content, highlighting a potential tradeoff between effectiveness and fairness. These findings advance the theoretical understanding of the functioning of different forms of moderation and are of high practical relevance for the moderation of online communities.

ACKNOWLEDGMENTS

We would like to thank the study participants for their time and efforts. We gratefully acknowledge support from the Institute for Ethics in Artificial Intelligence (IEAI) at the Technical University of Munich. We further thank the anonymous reviewers and editors for their insightful comments that helped improve the paper.

REFERENCES

- [1] Amalia Álvarez-Benjumea and Fabian Winter. 2018. Normative Change and Culture of Hate: An Experiment in Online Environments. *European Sociological Review* 34, 3 (2018), 223–237. <https://doi.org/10.1093/esr/jcy005>
- [2] Amnesty Global Insights. 2017. Unsocial Media: The Real Toll of Online Abuse Against Women. <https://medium.com/amnesty-insights/unsocial-media-the-real-toll-of-online-abuse-against-women-37134ddab3f4> Retrieved on March 30, 2021.
- [3] Reuben M. Baron and David A. Kenny. 1986. The Moderator–Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations. *Journal of Personality and Social Psychology* 51, 6 (1986), 1173–1182. <https://doi.org/10.1037/0022-3514.51.6.1173>
- [4] Roy F. Baumeister, Kathleen D. Vohs, and David C. Funder. 2007. Psychology as the Science of Self-Reports and Finger Movements: Whatever Happened to Actual Behavior? *Perspectives on Psychological Science* 2, 4 (2007), 396–403. <https://doi.org/10.1111/j.1745-6916.2007.00051.x>
- [5] Aparajita Bhandari, Marie Ozanne, Natalya N. Bazarova, and Dominic DiFranzo. 2021. Do You Care Who Flagged This Post? Effects of Moderator Visibility on Bystander Behavior. *Journal of Computer-Mediated Communication* 26, 5 (2021), 284–300. <https://doi.org/10.1093/jcmc/zmab007>

- [6] Christine Boomsma and Linda Steg. 2014. Feeling Safe in the Dark: Examining the Effect of Entrapment, Lighting Levels, and Gender on Feelings of Safety and Lighting Policy Acceptability. *Environment and Behavior* 46, 2 (2014), 193–212. <https://doi.org/10.1177/0013916512453838>
- [7] Bianca Cepollaro, Maxime Lepoutre, and Robert Mark Simpson. 2023. Counterspeech. *Philosophy Compass* 18, 1, Article e12890 (2023), 11 pages. <https://doi.org/10.1111/phc3.12890>
- [8] Robert B. Cialdini, Linda J. Demaine, Brad J. Sagarin, Daniel W. Barrett, Kelton Rhoads, and Patricia L. Winter. 2006. Managing Social Norms for Persuasive Impact. *Social Influence* 1, 1 (2006), 3–15. <https://doi.org/10.1080/15534510500181459>
- [9] Robert B. Cialdini, Carl A. Kallgren, and Raymond R. Reno. 1991. A Focus Theory of Normative Conduct: A Theoretical Refinement and Reevaluation of the Role of Norms in Human Behavior. In *Advances in Experimental Social Psychology*, Mark P. Zanna (Ed.). Vol. 24. Academic Press, 201–234. [https://doi.org/10.1016/S0065-2601\(08\)60330-5](https://doi.org/10.1016/S0065-2601(08)60330-5)
- [10] Robert B. Cialdini and Melanie R. Trost. 1998. Social Influence: Social Norms, Conformity and Compliance. In *The Handbook of Social Psychology*, Daniel T. Gilbert, Susan T. Fiske, and Gardner Lindzey (Eds.). McGraw-Hill, 151–192.
- [11] Christine L. Cook, Aashka Patel, and Donghee Yvette Wohn. 2021. Commercial Versus Volunteer: Comparing User Perceptions of Toxicity and Transparency in Content Moderation Across Social Media Platforms. *Frontiers in Human Dynamics* 3 (2021), 8 pages. <https://doi.org/10.3389/fhumd.2021.626409>
- [12] Niklas F. Cypriß, Severin Engelmann, Julia Sasse, Jens Grossklags, and Anna Baumert. 2022. *Intervening Against Online Hate Speech: A Case for Automated Counterspeech*. Research Brief. Institute for Ethics in Artificial Intelligence. https://ieai.sot.tum.de/wp-content/uploads/2022/05/Research-Brief_Intervening-Against-Online-Hate-Speech_April2022_FINAL.pdf Retrieved on March 30, 2023.
- [13] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G*Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences. *Behavior Research Methods* 39, 2 (2007), 175–191. <https://doi.org/10.3758/BF03193146>
- [14] Peter Glick and Susan T. Fiske. 1996. The Ambivalent Sexism Inventory: Differentiating Hostile and Benevolent Sexism. *Journal of Personality and Social Psychology* 70, 3 (1996), 491–512. <https://doi.org/10.1037/0022-3514.70.3.491>
- [15] Peter Glick and Susan T. Fiske. 2001. An Ambivalent Alliance: Hostile and Benevolent Sexism as Complementary Justifications for Gender Inequality. *American Psychologist* 56, 2 (2001), 109–118. <https://doi.org/10.1037/0003-066X.56.2.109>
- [16] João Gonçalves, Ina Weber, Gina M. Masullo, Marisa Torres da Silva, and Joep Hofhuis. 2021. Common Sense or Censorship: How Algorithmic Moderators and Message Type Influence Perceptions of Online Content Deletion. *New Media & Society* [Online First] (2021), 23 pages. <https://doi.org/10.1177/14614448211032310>
- [17] Jörg Gross and Alexander Vostroknutov. 2022. Why Do People Follow Social Norms? *Current Opinion in Psychology* 44 (2022), 1–6. <https://doi.org/10.1016/j.copsyc.2021.08.016>
- [18] Alon Halevy, Cristian Canton-Ferrer, Hao Ma, Umut Ozertem, Patrick Pantel, Marzieh Saeidi, Fabrizio Silvestri, and Ves Stoyanov. 2022. Preserving Integrity in Online Social Networks. *Communications of the ACM* 65, 2 (2022), 92–98. <https://doi.org/10.1145/3462671>
- [19] Nicola Henry and Anastasia Powell. 2018. Technology-Facilitated Sexual Violence: A Literature Review of Empirical Research. *Trauma, Violence, & Abuse* 19, 2 (2018), 195–208. <https://doi.org/10.1177/1524838016650189>
- [20] Jane Im, Sarita Schoenebeck, Marilyn Iriarte, Gabriel Grill, Daricia Wilkinson, Amna Batool, Rahaf Alharbi, Audrey Funwie, Tergel Gankhuu, Eric Gilbert, and Mustafa Naseem. 2022. Women’s Perspectives on Harm and Justice after Online Harassment. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2, Article 355 (2022), 23 pages. <https://doi.org/10.1145/3555775>
- [21] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 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, Article 192 (2019), 33 pages. <https://doi.org/10.1145/3359294>
- [22] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 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, Article 150 (2019), 27 pages. <https://doi.org/10.1145/3359252>
- [23] Jialun Aaron Jiang, Peipei Nie, Jed R. Brubaker, and Casey Fiesler. 2023. A Trade-off-Centered Framework of Content Moderation. *ACM Transactions on Computer-Human Interaction* 30, 1, Article 3 (2023), 34 pages. <https://doi.org/10.1145/3534929>
- [24] Matthew Katsaros, Tom Tyler, Jisu Kim, and Tracey Meares. 2022. Procedural Justice and Self Governance on Twitter: Unpacking the Experience of Rule Breaking on Twitter. *Journal of Online Trust and Safety* 1, 3, Article 38 (2022), 26 pages. <https://doi.org/10.54501/jots.v1i3.38>
- [25] Kees Keizer, Siegwart Lindenberg, and Linda Steg. 2008. The Spreading of Disorder. *Science* 322, 5908 (2008), 1681–1685. <https://doi.org/10.1126/science.1161405>

- [26] Tina Kuo, Alicia Hernani, and Jens Grossklags. 2023. The Unsung Heroes of Facebook Groups Moderation: A Case Study of Moderation Practices and Tools. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1, Article 97 (2023), 38 pages. <https://doi.org/10.1145/3579530>
- [27] Margaret Levi, Audrey Sacks, and Tom Tyler. 2009. Conceptualizing Legitimacy, Measuring Legitimizing Beliefs. *American Behavioral Scientist* 53, 3 (2009), 354–375. <https://doi.org/10.1177/0002764209338797>
- [28] D. Stephen Lindsay, Daniel J. Simons, and Scott O. Lilienfeld. 2016. Research Preregistration 101. *APS Observer* 29, 10 (2016), 7 pages. <https://www.psychologicalscience.org/observer/research-preregistration-101>
- [29] Binny Mathew, Anurag Illendula, Punyajoy Saha, Soumya Sarkar, Pawan Goyal, and Animesh Mukherjee. 2020. Hate Begets Hate: A Temporal Study of Hate Speech. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2, Article 92 (2020), 24 pages. <https://doi.org/10.1145/3415163>
- [30] Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. Thou Shalt Not Hate: Countering Online Hate Speech. In *Proceedings of the International AAAI Conference on Web and Social Media*. 369–380. <https://doi.org/10.1609/icwsm.v13i01.3237>
- [31] Jozef Miškolci, Lucia Kováčová, and Edita Rigová. 2020. Countering Hate Speech on Facebook: The Case of the Roma Minority in Slovakia. *Social Science Computer Review* 38, 2 (2020), 128–146. <https://doi.org/10.1177/0894439318791786>
- [32] Kevin Munger. 2017. Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment. *Political Behavior* 39 (2017), 629–649. <https://doi.org/10.1007/s11109-016-9373-5>
- [33] Marjan Nadim and Audun Fladmoe. 2021. Silencing Women? Gender and Online Harassment. *Social Science Computer Review* 39, 2 (2021), 245–258. <https://doi.org/10.1177/0894439319865518>
- [34] Daniel M. Oppenheimer, Tom Meyvis, and Nicolas Davidenko. 2009. Instructional Manipulation Checks: Detecting Satisficing to Increase Statistical Power. *Journal of Experimental Social Psychology* 45, 4 (2009), 867–872. <https://doi.org/10.1016/j.jesp.2009.03.009>
- [35] Marie Ozanne, Aparajita Bhandari, Natalya N. Bazarova, and Dominic DiFranzo. 2022. Shall AI Moderators Be Made Visible? Perception of Accountability and Trust in Moderation Systems on Social Media Platforms. *Big Data & Society* 9, 2, Article 20539517221115666 (2022), 13 pages. <https://doi.org/10.1177/20539517221115666>
- [36] Elizabeth Levy Paluck, Hana Shepherd, and Peter M. Aronow. 2016. Changing Climates of Conflict: A Social Network Experiment in 56 Schools. *Proceedings of the National Academy of Sciences* 113, 3 (2016), 566–571. <https://doi.org/10.1073/pnas.1514483113>
- [37] Kim Parker and Cary Funk. 2017. *Gender Discrimination Comes in Many Forms for Today's Working Women*. Survey Report. Pew Research Center. <https://www.pewresearch.org/short-reads/2017/12/14/gender-discrimination-comes-in-many-forms-for-todays-working-women/> Retrieved on January 10, 2023.
- [38] Jessica A. Pater, Moon K. Kim, Elizabeth D. Mynatt, and Casey Fiesler. 2016. Characterizations of Online Harassment: Comparing Policies Across Social Media Platforms. In *Proceedings of the 2016 ACM International Conference on Supporting Group Work*. 369–374. <https://doi.org/10.1145/2957276.2957297>
- [39] Kristopher J. Preacher and Andrew F. Hayes. 2004. SPSS and SAS Procedures for Estimating Indirect Effects in Simple Mediation Models. *Behavior Research Methods, Instruments, & Computers* 36 (2004), 717–731. <https://doi.org/10.3758/BF03206553>
- [40] Kristopher J. Preacher and Andrew F. Hayes. 2008. Contemporary Approaches to Assessing Mediation in Communication Research. In *The SAGE Sourcebook of Advanced Data Analysis Methods for Communication Research*, Andrew F. Hayes, Michael D. Slater, and Leslie B. Snyder (Eds.). Sage Publications, 13–54. <https://doi.org/10.4135/9781452272054.n2>
- [41] Punyajoy Saha, Kanishk Singh, Adarsh Kumar, Binny Mathew, and Animesh Mukherjee. 2022. CounterGeDi: A Controllable Approach to Generate Polite, Detoxified and Emotional Counterspeech. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*. 5157–5163. <https://doi.org/10.24963/ijcai.2022/716>
- [42] Julia Sasse, Niklas Cypri, and Anna Baumert. 2023. Online Moral Courage. In *Handbook of Peace Psychology*, Christopher Cohrs, Nadine Knab, and Gert Sommer (Eds.). Forum Friedenspsychologie e.V. <https://doi.org/10.17192/es2022.0074>
- [43] Julia Sasse, Jolien A. van Breen, Russell Spears, and Ernestine H. Gordijn. 2021. The Rocky Road from Experience to Expression of Emotions – Women's Anger About Sexism. *Affective Science* 2, 4 (2021), 414–426. <https://doi.org/10.1007/s42761-021-00081-7>
- [44] Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 111–125. <https://doi.org/10.1145/2998181.2998277>
- [45] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator Engagement and Community Development in the Age of Algorithms. *New Media & Society* 21, 7 (2019), 1417–1443. <https://doi.org/10.1177/1461444818821316>
- [46] Russell Spears and Tom Postmes. 2015. Group Identity, Social Influence, and Collective Action Online. In *The Handbook of the Psychology of Communication Technology*, S. Shyam Sundar (Ed.). John Wiley & Sons, Chapter 2, 23–46. <https://doi.org/10.1002/9781118426456.ch2>

- [47] Janet K. Swim and Laurie L. Cohen. 1997. Overt, Covert, And Subtle Sexism: A Comparison Between the Attitudes Toward Women and Modern Sexism Scales. *Psychology of Women Quarterly* 21, 1 (1997), 103–118. <https://doi.org/10.1111/j.1471-6402.1997.tb00103.x>
- [48] Janet K. Swim and Lauri L. Hyers. 1999. Excuse Me – What Did You Just Say?!: Women’s Public and Private Responses to Sexist Remarks. *Journal of Experimental Social Psychology* 35, 1 (1999), 68–88. <https://doi.org/10.1006/jesp.1998.1370>
- [49] Janet K. Swim and Lauri L. Hyers. 2009. Sexism. In *Handbook of Prejudice, Stereotyping, and Discrimination*, Todd D. Nelson (Ed.). Psychology Press, 407–430. <https://doi.org/10.4324/9781841697772>
- [50] Janet K. Swim, Lauri L. Hyers, Laurie L. Cohen, and Melissa J. Ferguson. 2001. Everyday Sexism: Evidence for Its Incidence, Nature, and Psychological Impact From Three Daily Diary Studies. *Journal of Social Issues* 57, 1 (2001), 31–53. <https://doi.org/10.1111/0022-4537.00200>
- [51] Tom Tyler, Matt Katsaros, Tracey Meares, and Sudhir Venkatesh. 2021. Social Media Governance: Can Social Media Companies Motivate Voluntary Rule Following Behavior Among Their Users? *Journal of Experimental Criminology* 17 (2021), 109–127. <https://doi.org/10.1007/s11292-019-09392-z>
- [52] Tom R. Tyler. 2006. Psychological Perspectives on Legitimacy and Legitimation. *Annual Review of Psychology* 57, 1 (2006), 375–400. <https://doi.org/10.1146/annurev.psych.57.102904.190038>
- [53] Adriane van der Wilk. 2018. *Cyber Violence and Hate Speech Online Against Women*. Study. Policy Department for Citizens’ Rights and Constitutional Affairs. [https://www.europarl.europa.eu/thinktank/en/document/IPOL_STU\(2018\)604979](https://www.europarl.europa.eu/thinktank/en/document/IPOL_STU(2018)604979)
- [54] Jessica Vitak, Kalyani Chadha, Linda Steiner, and Zahra Ashktorab. 2017. Identifying Women’s Experiences With and Strategies for Mitigating Negative Effects of Online Harassment. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 1231–1245. <https://doi.org/10.1145/2998181.2998337>
- [55] Emily Vogels. 2021. *The State of Online Harassment*. Survey Report. Pew Research Center. <https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/> Retrieved on March 30, 2021.
- [56] Wanzheng Zhu and Suma Bhat. 2021. Generate, Prune, Select: A Pipeline for Counterspeech Generation against Online Hate Speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 134–149. <https://doi.org/10.18653/v1/2021.findings-acl.12>

Received January 2023; revised April 2023; accepted May 2023