Bystanders of Online Moderation: Examining the Effects of Witnessing Post-Removal Explanations

Shagun Jhaver
Rutgers University
New Brunswick, NJ, USA
shagun.jhaver@rutgers.edu

Himanshu Rathi
Rutgers University
New Brunswick, NJ, USA
hr393@scarletmail.rutgers.edu

Koustuv Saha
University of Illinois
Urbana-Champaign
Urbana, IL, USA
ksaha2@illinois.edu

ABSTRACT

Prior research on transparency in content moderation has demonstrated the benefits of offering post-removal explanations to sanctioned users. In this paper, we examine whether the influence of such explanations transcends those who are moderated to the bystanders who witness such explanations. We conduct a quasi-experimental study on two popular Reddit communities (*r/AskReddit* and *r/science*) by collecting their data spanning 13 months—a total of 85.5M posts made by 5.9M users. Our causal-inference analyses show that bystanders significantly increase their posting activity and interactivity levels as compared to their matched control set of users. In line with previous applications of Deterrence Theory on digital platforms, our findings highlight that understanding the rationales behind sanctions on another user significantly shapes observers' behaviors. We discuss the theoretical implications and design recommendations of this research, focusing on how investing more efforts in post-removal explanations can help build thriving online communities.

CCS CONCEPTS

• **Human-centered computing** → *Empirical studies in collaborative and social computing*; *Social media.*

KEYWORDS

content moderation, social media, transparency, causal-inference

ACM Reference Format:
Shagun Jhaver, Himanshu Rathi, and Koustuv Saha. 2024. Bystanders of Online Moderation: Examining the Effects of Witnessing Post-Removal Explanations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3613904.3642204

1 INTRODUCTION

As the social media ecosystem continues to rapidly expand, platform designers and researchers are experimenting with new models of digital governance [24, 40, 59]. Recent research has begun extending guiding principles that could possibly serve such models [60, 75]. This includes rights-based legal approaches, such as international human rights law and American civil rights law [14]. The HCI community has especially centered around aspirational computer science principles of fairness, accountability, transparency, ethics, and responsibility [71]. Most famously, a group of human rights organizations, advocates, and academic experts developed and launched what they termed "the Santa Clara Principles on Transparency and Accountability in Content Moderation," which aim to guide platforms on how to incorporate meaningful transparency and accountability around moderation of user-generated content [58].

Again, empirical research on incorporating meaningful transparency and how it may benefit users as well as platforms, has begun to emerge. For example, transparency through removal notification and providing moderators' reasoning behind content removal has been shown as one of the key factors in users' perception of the fairness of content moderation [19]. Another study has shown that when offered removal explanations in any online community, users tend to improve their posting behavior in that community in the future [22]. Such evidence has been used to motivate platforms, community moderators, and policymakers to continue to push for increased, meaningful transparency in their moderation practices.

This study seeks to add further empirical evidence to the effects of offering transparency in content moderation on social media platforms. Specifically, we look at whether such transparency can serve users other than those sanctioned. Prior research has provided evidence for the educational benefits of offering removal explanations for users whose content is removed [19, 22]. However, the effects on *bystanders* who witness the post-removal and the explanation behind it have not been tested. Focusing on bystanders allows us to examine the impact of indirect experiences with punishment on users' behavior. In this research, we ask the question: *Do public removal explanations intended for the sanctioned users influence the posting behavior of bystanders to those explanations?*

We collected a dataset of 85.5M posts from two large Reddit communities, *r/AskReddit* and *r/science*, over the time period Dec 2021–Dec 2022, and developed a computational framework based on causal inference that matched users who witnessed a removal explanation in June 2022 with users who did not witness any explanation. Comparing the post-treatment behavior of these matched groups, we found that exposure to removal explanations significantly boosted the posting activity and interactivity of bystanders as compared to non-bystanders. This shows that the behavioral impacts of moderation transparency on posting volumes are more broadly applicable than previously understood [19, 22]. Drawing upon this insight, we argue that community managers must invest more time and effort in increasing moderation transparency

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05…$15.00
https://doi.org/10.1145/3613904.3642204

through explanation messages. On the other hand, witnessing explanation messages did not significantly enhance the posting quality of bystanders. We speculate on the causes of this empirical insight and offer directions for future research that may help us better understand the role of explanation messages.

2 BACKGROUND AND RELATED WORK

2.1 Transparency in Content Moderation

Moderation systems on social media platforms are designed for governance purposes and often impose measures such as removing content, muting, or banning offenders [11, 15]. These measures are implemented by content moderators, who may either be volunteers among the platform's user base or commercial content moderators hired by the platform [49, 63]. More recently, AI-driven tools have been used to assist in moderation processes [1, 16, 27, 32]. We focus here on transparency in end-users' experience with moderation processes. *Transparency* implies opening up "the working procedures not immediately visible to those not directly involved to demonstrate the good working of an institution" [46].

We situate our work within a line of research that examines the impact of content moderation on end-users. Scholars have investigated the impact of both user-level [19, 21, 62, 72] and community-wide sanctions [5, 6]. This has included studies using a variety of methods, such as interviews [27], design workshops [71], surveys [19, 73], and log analyses [5, 6, 21]. Prior work has also highlighted the benefits of offering moderation explanations to sanctioned users [19, 22]. We focus on end-users who witness, although not directly affected by, the moderation sanctions. By doing so, we contribute to building a theory [33] that prescribes to community managers which moderation interventions should be deployed, under what circumstances, and with what expected outcomes.

In examining the complexities of enacting content moderation, researchers have identified several issues regarding transparency in the procedures followed by platforms when applying punitive measures [39]. First, the criteria of inappropriate content might not be well-established before moderation decisions are made [61]. Legal experts have raised concerns that despite social media platforms publicly sharing their content policies, they often fail to adequately consider the contextual factors surrounding the content, such as its localized meaning and the identities of the speakers and audiences when evaluating its appropriateness [74]. Second, there are inter-platform differences in how norm violations are conceptualized. For example, an HCI study comparing the content policies of 15 platforms found a lack of consensus in defining what qualifies as online harassment and how forcefully content deemed as harassment should be moderated [48]. Consequently, when these vague content policies are implemented for content regulation, it can cause ambiguity in resolving moderation cases [74]. Finally, and most pertinent to our study, communication with end-users on moderation decisions is often found to be deficient in details [66, 73].

2.2 Removal Explanations and Bystanders to Norm Violations

Prior research has emphasized the significance of incorporating moderation notifications and explanations into the design of moderation systems [22, 35, 37, 70]. For example, researchers have shown that when Facebook and Reddit platforms do not inform users about their content removal [66], users question which platform policy they have violated [19, 73]. Besides removal notification, users desire a justification for why their posts got removed, deeming it a significant factor in their perception of moderation fairness [19]. Users also express dissatisfaction with the inconsistent punishments meted out to them versus others, leading them to request explanations further [38, 70]. Many studies have empirically shown the benefits of offering removal explanations in improving the behavior of moderated users [19, 22, 69]. For example, Tyler et al. found that users who were provided education about platform rules in the week following their post removal were less likely to post new violating content [69]. We extend this research by investigating the utility of explanations in influencing the behavior of bystanders.

Curiously, Reddit moderators offer explanations publicly by commenting on the removed submission. While this is not the sole communication mode—indeed, many moderators privately message users to inform them about moderation [22, 53]—prior research has argued that public explanations serve to enhance broader transparency efforts [19, 22]. On Reddit, users already engaging with a post retain access to it even after it is removed from the main subreddit; in this sense, removed submissions are not really *removed*, just hidden from the public view. By publicly explaining the reason behind post removal, explanation comments serve users who stumble upon it or are already engaged.

We extend prior inquiries into using Deterrence Theory [65] to evaluate the impact of punishments on deterring inappropriate behaviors online [12, 62]. Deterrence Theory makes a distinction between general and specific deterrence—specific deterrence refers to the effect of punitive measures on individuals subjected to them, while general deterrence pertains to the impact of the potential threat of such measures on uninvolved observers. By focusing on bystanders, we examine the effects of generalized deterrence in shaping user behavior. Seering et al. showed that banning any type of behavior on Twitch significantly reduced the frequency of that behavior in subsequent messages posted by others [62]. Building upon this, we examine whether clarifying which aspects of submissions prompt sanctions via explanation messages influences observers' subsequent actions.

Encouraging voluntary compliance with behavioral norms in a community requires that community members know the norms and be aware of them when being active within the community. Kiesler et al. [33] argue that people learn the community norms in three ways: (1) observing other people's behavior and its consequences, (2) seeing codes of conduct, and (3) behaving and directly receiving feedback. Prior research has demonstrated the importance of users seeing codes of conduct [41] and directly receiving feedback in improving their subsequent behavior [22, 69]. We focus here on establishing the utility of bystanders observing other people's norm violations and the resulting consequences.

In terms of reducing the posting of norm-violating content, some research has focused on the roles bystanders can play in the context of online harassment. Blackwell et al. found that labeling a variety of technology-enabled abusive experiences as 'online harassment' helps bystanders *understand* the breadth and depth of this problem [3]. Further, designs that motivate bystander intervention discourage harassment through normative enforcement [2]. Taylor

et al. [67] additionally found that design solutions that encourage empathy and accountability can promote bystander intervention in cyberbullying. Extending this line of research to a broader range of norm violations, we analyze how bystanders are affected by their exposure to post-removal explanations.

3 DATA AND METHODS

3.1 Study Design and Rationale

We conducted an observational study to examine the effects of witnessing post-removal explanations on Reddit. Prior HCI and CSCW research has recognized that observational analyses of social media data can serve as a valuable tool for understanding society and evaluating changes in users' behavior, especially regarding their use of social network sites [64]. Regarding our study's context, empirical research on the effects of various content moderation interventions has often deployed observational analyses of social media logs [5, 6, 52, 62]. Similar to our work, such research has primarily examined behavior patterns over more extended timeframes, typically spanning months [17, 21, 22].

Examining the impact of an intervention, whether internal or external, is best studied through causal inference approaches, such as randomized controlled trials (RCTs). However, these approaches have certain limitations. First, experimental studies requiring participant consent can be constrained by concerns about the observer effect [43]—that individuals might alter their typical behavior when they are aware of being monitored or observed. Second, conducting experimental research without participants' awareness is considered unethical, especially within the human-centered research paradigm [29, 45]. Finally, conducting experiments without prior awareness of their potential impact on participants can lead to long-term adverse consequences for both platforms and individuals.

As a result, observational studies can serve as a viable alternative in situations where experimental approaches may not be feasible or ethical. While observational studies may not provide true causality, they are structured to minimize confounds and investigate longitudinal data, offering stronger evidence than basic correlational analyses [18]. Recently, there has been growing interest in these types of studies within the fields of HCI and behavioral science, including those analyzing social media data [8, 30, 31, 47, 51, 54, 56, 77]. Significantly, the research conducted by Saha et al. prompted us to operationalize metrics for assessing social media behavior, including factors like activity and interactivity [57].

Given the above considerations, we drew on quasi-experimental approaches to observational data. We adopted a causal-inference approach based on the potential outcomes framework proposed by Rubin [50]. Fig. 2 shows a schematic figure of our approach. This approach simulates an experimental setting by matching individuals (Treated and Control) on several covariates [18]. For a given treatment, T, two potential outcomes are compared: (1) when a user is exposed to T ($T = 1$), and (2) when a user is not exposed to T ($T = 0$). Because it is impossible to obtain both kinds of outcomes simultaneously for the same user, this framework estimates the missing counterfactual for a user based on the outcomes of a matched user—another user with similar covariates (attributes and behaviors) but not exposed to T. Our work drew motivation from prior works that adopted similar causal-inference approaches on social media data [6, 31, 52, 56].

Table 1: Summary statistics of the Reddit dataset.

Subreddit	No. Submissions	No. Comments
r/Askreddit	287,954	5,358,662
r/science	2,453	175,007

3.2 Choice of Subreddits

This paper focuses on two major subreddits, *r/AskReddit* (43M members) and *r/science* (31M members). r/AskReddit is a community focused on asking and answering questions that elicit thought-provoking discussions, offer light entertainment, and help users learn more about their fellow community members.[1] r/science is a science news and discussion community where users post links to research papers or reputable news items representing recent scientific research, and engage in science communication [28].

We analyze these two communities for two main reasons. First, due to their importance—they are among the largest and most active Reddit communities and have impacted society at large, e.g., through widespread sharing of personal experiences, expert testimony, and science communication on a range of topics [28, 36]. Second, both communities have a mature moderation approach—they have been active for more than 15 years, and have a well-described set of posting guidelines and dozens of active moderators. This made it more likely that their approach to offering removal explanations would deliver messages appropriate for our study.

Fig. 1 shows example post-removals on the subreddits. We downloaded the data from these subreddits over 13 months between 01 December 2021–31 December 2022, using the *pushshift.io* service.

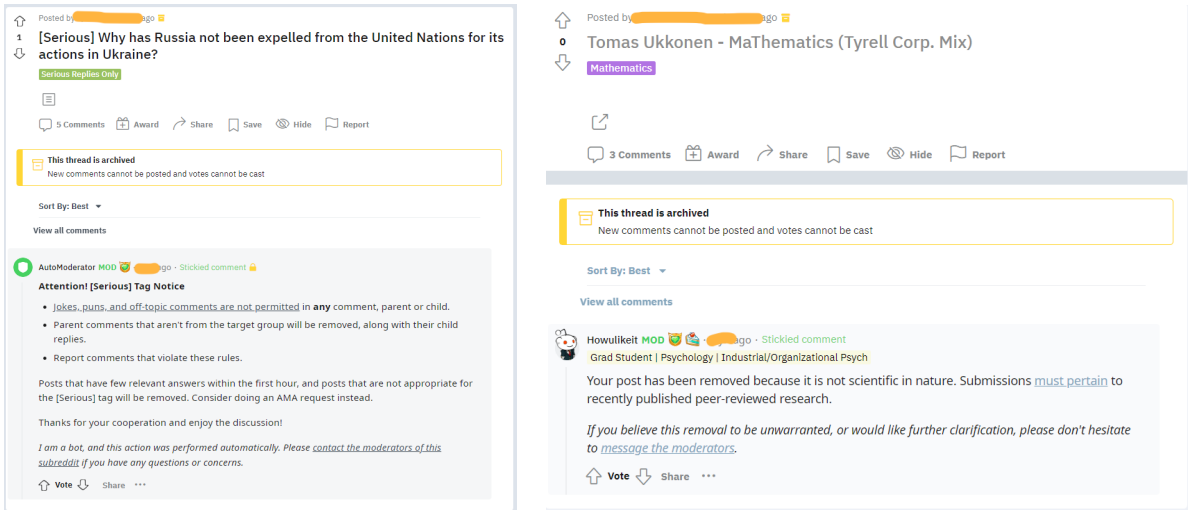
We iterated through this dataset, decompressing and decoding it in smaller chunks, and simultaneously storing the readable data into SQLite database tables. We queried the database to access the data for the ensuing analyses in the paper. Table 1 summarizes the data (submissions and comments) collected for our study. Note that we use the term *post* to indicate posting activity in the form of either submissions or comments; therefore, for any given period T, $\mathsf{N_p(T) = N_s(T) + N_c(T)}$ (where $\mathsf{N_p}$, $\mathsf{N_s}$, and $\mathsf{N_c}$ denote the number of posts, submissions, and comments respectively).

3.3 Defining Treated and Control Users

Our study employed a causal-inference framework, drawing on similar approaches in prior research [6, 31, 57]. For this purpose, we defined *treatment* as exposure to post-removal explanation(s). Within our study period of 13 months, we considered the period between 01-30 June 2022, as our *treatment* period, i.e., we focused on explanations provided in this one month and collected six-month pre-treatment and six-month post-treatment period data for our analyses. We randomly selected June 2022 as our treatment period, following similar selections in prior moderation research [22].

Our Treated users comprise the "bystanders" or the users of a subreddit who witnessed a removal explanation during the treatment period. While this set would ideally consist of users who

[1] https://www.reddit.com/r/AskReddit/wiki/index



(a) (b)

Figure 1: Examples of post-removals and explanations by a moderator on (a) *r/AskReddit* (here, the explanation is provided by the Automoderator), and (b) *r/science* (here, the explanation is provided by a human moderator).

read and comprehended the explanation comment, we did not have access to users' viewing logs. Therefore, we constituted Treated users by assuming the commenting activity as a proxy for exposure and including users who commented in the discussion thread containing the removal explanation. On the other hand, Control users comprise users of the same subreddit who did not comment in any discussion thread containing a removal explanation but posted elsewhere in the same subreddit during the pre-treatment period. We filtered out the data of any user exposed to post-removal explanations in the period between December 2021–May 2022 to ensure that we examined Treated users subjected to *treatment* in June 2022.

3.4 Gathering Post Removal Explanations Data

We obtained a list of 95 phrases indicating post-removal explanations from prior work by Jhaver et al. [22]. We used these phrases to query our database to collect all the removal explanations in our defined *treatment* period. Specifically, we queried the created database to retrieve the stickied[2] comments made by moderators within the *treatment* period and containing any of the above phrases. We obtained 257 removal explanations on *r/AskReddit* and 379 such removal explanations on *r/science*. Focusing on the discussion threads of each of these removal explanations, we next collected the information of the commenters, who are the bystanders or Treated users in our study. In some threads, the removed submission's author also posted a comment in the discussion thread; we did not include such submission authors in the Treated users groups because our analysis centers on bystanders, not moderated users.

We obtained the timeline of posts made by the Treated users in the corresponding subreddit during the study period. For each subreddit, we also curated a list of Control users: this constituted users who were not Treated users, and who were not exposed to

[2]Removal explanations are usually stickied, i.e., locked to appear as the top comment in the discussion thread.

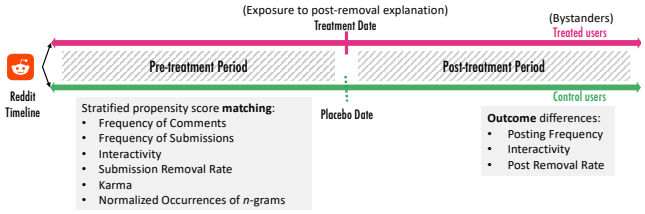


Figure 2: A schematic figure showing our causal-inference approach to analyze users' Reddit timeline.

any post-removal explanations in the pre-treatment period. The Treated users were assigned with a *treatment date* on their first occurrence of witnessing a post-removal explanation during our *treatment period*. On the other hand, because the Control users did not have any *treatment date* per se, we simulated a set of *placebo dates* from the set of all possible treatment dates within the subreddit, such that the distributions of *placebo dates* and *treatment dates* were statistically similar. Then, each Control user was randomly assigned a *placebo date* from the set of *placebo dates*. For easier readability, any following reference to pre-treatment and post-treatment surrounds *treatment date* for a Treated user, and *placebo date* for a Control user.

3.5 Matching for Causal-Inference

3.5.1 Covariates for Matching. We operationalized a number of covariates that we would use for matching the Treated and Control users, motivated from prior work [6, 9, 22, 31, 55, 57], as listed below. Each covariate was measured using the data in the user's pre-treatment history.

–*Frequency of Comments* : The normalized quantity of comments per day, as also used in prior work [6, 57].

–*Frequency of Submissions* : The normalized quantity of submissions per day, as also used in prior work [6, 57].

–*User Interactivity* : The ratio of number of comments to the total number of posts, as also used in prior work [55, 57].

–*Submission Removal Rate* : The ratio of removal submissions to total submissions posted by the user, as also used in prior work [22].

–*Karma* : Average karma across the comments and submissions made by the user, as also used in prior work [6, 57].

–*Normalized n-grams* : The normalized occurrences of the top 1000 n-grams ($n = 1, 2$),as also used in prior work [9, 56].

3.5.2 Stratified Propensity Score Matching. As mentioned above, we used matching to find pairs (generalizable to groups) of Treated and Control users with statistically similar covariates. We adopted the propensity score matching approach that matches users based on propensity scores, which is essentially a user's *likelihood* of receiving the treatment. However, exact one-to-one propensity score matching can suffer from biases [34]. Therefore, motivated by prior work [31, 54, 76], we adopted *stratified propensity score matching* that can balance the bias-variance tradeoff of either too biased (one-to-one match) or too variant (unmatched) data comparisons. In a stratified matching approach, users with similar propensity scores are grouped into strata. Hence, every stratum consists of users with similar covariates [31]. Through this approach, we isolated and estimated treatment effects within each stratum.

For the above matching, we computed the propensity scores by building a logistic regression model with the covariates as independent variables and a user's binary treatment score (1 for Treated users and 0 for Control users) as dependent variable. We segregated the distribution of propensity scores into 200 strata of equal width. To ensure that our causal analysis was restricted to a sufficient number of similar users, we discarded strata with less than 10 Treated and 10 Control users. This led to a final matched dataset of 50 strata (4,842 Treated users and 146,922 Control users) in *r/AskReddit* and 33 strata (4,890 Treated users and 176,324 Control users) in *r/science*.

3.6 Measuring Treatment Effects

After matching the Treated and Control users, we measured the differences in the post-treatment behaviors of the users. For this, we operationalized three outcomes—1) *Frequency of posting*, 2) *Interactivity*, and 3) *Submission Removal Rate* for the users in the post-treatment period. We draw on the difference in differences approach in causal-inference [13] and prior work using these approaches on social media [6, 9, 55, 56], to calculate the average treatment effect (ATE) as the average of the difference of changes in the Treated users and the Control users per stratum. In addition, we obtained the effect size (Cohen's d) and evaluated statistical significance in differences using relative t-tests. We conducted Kolmogorov-Smirnov (KS) test to evaluate the differences in the distributions of the Treated and Control groups' outcomes.

4 RESULTS

Table 2 summarizes our observations of the differences in the post-treatment outcomes in our study. We describe our findings below:

Posting Frequency. We find significant differences in the posting frequency of Treated and matched Control individuals. On *r/AskReddit*, the ATE is 0.453, which can be roughly interpreted as the treatment increases the frequency of posts by 1 for about 45.3% of the individuals. We see a high effect size (0.807) and significant differences as per t-test and KS-test ($p < 0.0001$). We also see convergent findings in *r/science* with an ATE of 0.025, Cohen's d of 1.075, and significant differences as per t-test and KS-test ($p < 0.0001$). Higher posting frequency indicates that the Treated users (bystanders) became more active in the subreddits after witnessing the post-removal explanations. This measure is an indicator of positive community behavior [57].

Table 2: Summary of changes in outcomes for the Treated and Control individuals. We report average treatment effect (ATE), effect size (Cohen's d), relative t-test, and KS-test statistics (* $p < 0.01$, ** $p < 0.001$, *** $p < 0.0001$).

Outcome	ATE	Cohen's d	t-test	KS-test
r/AskReddit				
Posting Frequency	0.453	0.807	6.589***	0.640***
Interactivity	0.193	2.392	12.233***	0.960***
Post Removal Rate	0.000	0.005	0.024	0.200
r/science				
Posting Frequency	0.025	1.075	8.890***	0.515***
Interactivity	0.216	1.445	17.469***	0.879***
Post Removal Rate	0.001	0.007	0.177	0.303

Interactivity. Similar to the above, we find significant differences in the interactivity of Treated and Control individuals. On *r/AskReddit* we find an ATE of 0.193 and Cohen's d of 2.392, along with statistical significance in differences as per t-test and KS-test (p<0.0001). Likewise, on *r/science*, ATE on interactivity is 0.216, Cohen's d is 1.445, and t-test and KS-tests reveal statistical significance (p<0.0001). In addition to higher posting frequency, higher interactivity indicates that the Treated users not only created more new submissions but also replied more to others' threads—an important factor for enhancing online community engagement [55, 57]. This suggests that post-removal explanations can potentially enhance community engagement and, subsequently, the sustainability and growth of a community with member activity.

Post Removal Rate. Interestingly, we find no significant effects on the post-removal rates. That is, we do not have conclusive evidence if the posting quality improved (or worsened) for the Treated users.

5 DISCUSSION

5.1 Implications

Online communities rely on content generated by users, but inappropriate posts can detract from the quality of the user experience. Consequently, moderation systems typically aim to boost the overall volume of contributions while reducing the need for post-removals [15, 33]. Our analysis in this paper examined the behavioral impact of offering moderation explanations on bystanders over two dimensions—their future posting activity and the frequency of their future post-removals. Our results represent the impact of generalized deterrence — the indirect experience with punishment. Consistent with prior applications of Deterrence Theory in online platforms [12, 62], we show that understanding the reasons for

sanctions on another user significantly shapes observers' behaviors. In this section, we examine the implications of our findings for moderators, site administrators, designers, and future research.

5.1.1 Removal Explanations Help Boost Posting Frequency. We found that on both *r/AskReddit* and *r/science*, users who got exposed to removal explanations directed at moderated others significantly increased their posting activity as compared to users who did not witness any explanations. It could be that seeing explanation messages indicated to bystanders that the community is well-moderated. This, in turn, could have enhanced their inclination to be active within the community.

We note that this result contrasts Jhaver et al.' findings for moderated users—exposure to removal explanations reduced these users' future posting activity [22]. One reason for this could be that users who suffer moderation may find it more difficult to accept the justification for their post-removals than other bystanders. Prior work has often grappled with the tradeoffs of moderation actions reducing posting traffic at the cost of improving posting quality [17, 19, 21, 22]. However, as this study's framing highlights, for any given removed submission, there is only one moderated user but potentially many more bystanders. Thus, our results suggest that providing explanation messages may boost the overall posting frequency in a community. This empirical insight offers a powerful incentive to community managers considering the deployment of explanation messages.

5.1.2 Removal explanations help increase community engagement. We found that exposure to others' explanation messages increases the posting interactivity. That is, bystanders' comments constitute a greater proportion of their posting volume after the treatment. Prior research has shown that this metric is an important factor in community engagement [57]. Therefore, this finding suggests that observing the reasoned explanation for post removals can inform bystanders why certain types of posts are unacceptable in the community, help them learn its accepted norms [7], and thereby increase their confidence in instituting a deeper engagement with the community. This further demonstrates the utility of offering post-removal explanations.

Another explanation for this finding is that users perceive moderators attend to and regulate inappropriate submissions more than inappropriate comments. This perception may incline them to engage more in posting comments than submissions to avoid experiencing post removals. As prior research shows, users often develop "folk theories" of content moderation processes to make sense of them [10, 19]. Going forward, qualitative studies could inquire whether the posting activity of users is shaped by their folk theories of where the content moderation efforts are focused.

5.1.3 Removal explanations do not impact post removals. Our analysis shows that removal explanations do not significantly impact the future post-removals of bystanders. This contrasts previous results for moderated users: Jhaver et al. showed that offering removal explanations reduced the future post-removals of moderated users [22]. This suggests that explanation messages boost the posting *quality* of moderated users more than bystanders. Why is this the case? One reason could be that having experienced post removal, moderated users may be likelier to attend to *all* community guidelines before posting their next submissions. On the other hand, witnessing a removal explanation may not be a strong enough incentive for bystanders to ensure compliance with all community guidelines in their next submissions.

It is possible that witnessing explanation messages educates bystanders about the violated community norm specific to the corresponding removed post and leads them to avoid the same violation in the future, yet they continue violating other community norms. While beyond the scope of the current paper, a more granular analysis could examine whether norm-specific learning occurs through removal explanations among bystanders. Besides, prior research has shown that users often respond to moderation by changing their deviant posting activities to circumvent restrictions, especially when the moderation is automated and reliant on detecting specific keywords [4, 23]. Thus, removal explanations may offer users clues on how to avoid moderation, thereby depicting a paradox of enacting algorithmic transparency [26]. Therefore, beyond focusing on post removals, it is important to qualitatively evaluate the extent to which removal explanations prompt users to *sincerely* engage in adhering to the community's expectations.

5.1.4 Design Implications. This work bears design implications regarding the positive impacts of enacting transparency in online content moderation. The empirical evidence presented here informs community managers to put more effort into providing explanations for sanctions, and more importantly, make these explanations *publicly visible*, so that they can educate bystanders. While content moderation actions have proliferated to align with the growing scale of online communities, providing explanations is still not as prevalent. For instance, to conduct this study, we originally started with four large subreddits—we had also collected over ~2M posts from *r/politics* (8.4M users) and *r/technology* (15M users). However, despite being large subreddits with many moderators, neither of these communities provided any post-removal explanations (which also prevented us from including their data in our analyses). Prior work has noted challenges in providing explanations in all instances, such as moderator fatigue and limitations of automated moderation tools [20, 22]. Many platforms may lack resources to provide moderation explanations. However, with the advent of generative AI and large-language model-based technologies, it would be interesting to explore the design space of curating automated explanation messages through these emerging technologies. Given that user attention is a limited resource [33], platforms must also negotiate the extent to which norm education through removal explanations intended for others be prioritized in the content shown to the users.

Besides, more research is needed to develop best practices for designing removal explanations in response to specific norm violations and other contextual details. The computational framework of our study can be extended to delineate the effects of different features of explanation messages, e.g., explanation length, politeness level, clarifying future graduated sanctions, including face-saving mechanisms [33]. The results of such analyses can inform platform owners and community managers about the suitability of different explanation types. Given the inherent connection of explanations to community guidelines, these efforts could also inform the latter's design. On Reddit, explanations are made publicly visible through

a stickied comment on the removed post. The visibility of such explanations can be further enhanced by sending notifications about them to users engaged in the sanctioned discussion threads.

5.2 Limitations and Future Directions

Our study focused on two large subreddits, so, our results are most readily applicable to other subreddits of similar size. Future analyses would benefit from investigating the circumstances under which these results replicate (or do not) on other platforms and communities. The computational framework we have presented here should help such inquiries. Prior similar efforts on developing extendable computational frameworks for evaluating moderation actions have similarly used data from a limited number of samples [5, 6, 21, 68].

For this work, we initially planned a comparative analysis of the effects of human v/s bot explanations on bystanders. However, our data showed that all *r/AskReddit* explanations were provided by bots and all *r/science* explanations by human moderators during the treatment period. Therefore, we could not conduct our planned comparative analysis for either community. Future work should explore how AI-generated explanations compare to human-offered explanations in influencing bystanders' behavior, extending similar inquiries in prior research [22]. Additionally, it would be fruitful to investigate how explanation messages shape other aspects of bystanders' behavior, e.g., their use of language and how other community members respond to them.

Our analysis does not consider the in-situ practical concerns and constraints under which content moderators work [42, 44]. Therefore, studies that examine how moderators draft, choose, and submit explanation messages, and help create tools that can make the workflow easier would empower moderators to send explanation messages at a higher frequency.

Our data collection constitutes Treated users by including everyone who commented in the discussion thread regardless of when they commented vis-a-vis the explanation message timestamp. This choice was inspired by our observation that Reddit users can access their posting history and may track the discussion long after their comment, especially since Reddit sends users notifications about posting activity in the threads where users contribute. Since explanation comments are highlighted at the top of the thread regardless of upvotes and posting time, exposure to them is likely for everyone viewing the thread. Still, some users might leave the discussion before the explanation message was posted and never returned. Further, some might have stumbled upon the thread and viewed the explanation message but never commented on the thread. Therefore, our measure of exposure to explanations is limited by our data access. Future research can measure this exposure more precisely by tracking users' passive consumption of explanation messages.

6 CONCLUSION

Transparency in communications is a key concern for moderated users [23, 25, 66]. On the other hand, secretiveness about moderation decisions triggers speculation among users who suspect potential biases [19, 23, 73]. In this paper, we focus on one important mode of enacting greater transparency in moderation decisions: publicly visible messaging by moderators that reveals the reasons behind submission removals. Our analysis shows that witnessing such messages significantly boosts the posting and interactivity levels of bystanders. This suggests that adopting an educational approach to content moderation, as opposed to a strictly punitive one, can lead to enhanced community outcomes.

ACKNOWLEDGMENTS

We thank Anish Gupta, Gayathri Ravipati, and Navreet Kaur for their work on this project.

REFERENCES

[1] Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In *International Conference on Social Informatics*. Springer, 405–415.
[2] Lindsay Blackwell, Tianying Chen, Sarita Schoenebeck, and Cliff Lampe. 2018. When Online Harassment is Perceived as Justified. In *Twelfth International AAAI Conference on Web and Social Media*.
[3] Lindsay Blackwell, Jill P Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. *PACMHCI* 1, CSCW (2017), 24–1.
[4] Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. 2016. #thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing* (San Francisco, California, USA) *(CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 1201–1213. https://doi.org/10.1145/2818048.2819963
[5] Eshwar Chandrasekharan, Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2022. Quarantined! Examining the Effects of a Community-Wide Moderation Intervention on Reddit. *ACM Trans. Comput.-Hum. Interact.* 29, 4, Article 29 (mar 2022), 26 pages. https://doi.org/10.1145/3490499
[6] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on human-computer interaction* 1, CSCW (2017), 1–22.
[7] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 32.
[8] Munmun De Choudhury, Shagun Jhaver, Benjamin Sugar, and Ingmar Weber. 2016. Social Media Participation in an Activist Movement for Racial Equality. In *Tenth International AAAI Conference on Web and Social Media* (2016).
[9] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2098–2110.
[10] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. I always assumed that I wasn't really that close to [her]: Reasoning about Invisible Algorithms in News Feeds. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. ACM, 153–162.
[11] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.
[12] Rosalie Gillett, Joanne E. Gray, and D. Bondy Valdovinos Kaye. 2023. 'Just a little hack': Investigating cultures of content moderation circumvention by Facebook users. *New Media & Society* 0, 0 (2023), 14614448221147661. https://doi.org/10.1177/14614448221147661 arXiv:https://doi.org/10.1177/14614448221147661
[13] Andrew Goodman-Bacon. 2021. Difference-in-differences with variation in treatment timing. *Journal of Econometrics* 225, 2 (2021), 254–277.
[14] Robert Gorwa. 2019. What is platform governance? *Information, Communication & Society* (2019), 1–18.
[15] James Grimmelmann. 2015. The virtues of moderation. *Yale JL & Tech.* 17 (2015).
[16] Manoel Horta Ribeiro, Justin Cheng, and Robert West. 2023. Automated Content Moderation Increases Adherence to Community Guidelines. In *Proceedings of the ACM Web Conference 2023*. 2666–2676.
[17] Manoel Horta Ribeiro, Shagun Jhaver, Savvas Zannettou, Jeremy Blackburn, Gianluca Stringhini, Emiliano De Cristofaro, and Robert West. 2021. Do Platform Migrations Compromise Content Moderation? Evidence from r/The_Donald and r/Incels. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 316 (oct 2021), 24 pages. https://doi.org/10.1145/3476057
[18] Guido W Imbens and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.
[19] 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. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 192 (Nov. 2019), 33 pages. https://doi.org/10.1145/3359294

[20] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Trans. Comput.-Hum. Interact.* 26, 5, Article 31 (July 2019), 35 pages. https://doi.org/10.1145/3338243

[21] Shagun Jhaver, Christian Boylston, Diyi Yang, and Amy Bruckman. 2021. Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 381 (oct 2021), 30 pages. https://doi.org/10.1145/3479525

[22] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does Transparency in Moderation Really Matter? User Behavior After Content Removal Explanations on Reddit. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 150 (nov 2019), 27 pages. https://doi.org/10.1145/3359252

[23] Shagun Jhaver, Quan Ze Chen, Detlef Knauss, and Amy X. Zhang. 2022. Designing Word Filter Tools for Creator-Led Comment Moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 205, 21 pages. https://doi.org/10.1145/3491102.3517505

[24] Shagun Jhaver, Seth Frey, and Amy X. Zhang. 2023. Decentralizing Platform Power: A Design Space of Multi-level Governance in Online Social Platforms. *Social Media + Society* 9, 4 (2023), 1–12. https://doi.org/10.1177/20563051231207857

[25] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online Harassment and Content Moderation: The Case of Blocklists. *ACM Trans. Comput.-Hum. Interact.* 25, 2, Article 12 (March 2018), 33 pages. https://doi.org/10.1145/3185593

[26] Shagun Jhaver, Yoni Karpfen, and Judd Antin. 2018. Algorithmic Anxiety and Coping Strategies of Airbnb Hosts. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (, Montreal QC, Canada,) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173995

[27] Shagun Jhaver, Alice Qian Zhang, Quan Ze Chan, Nikhila Natarajan, Ruotong Wang, and Amy X. Zhang. 2023. Personalizing Content Moderation on Social Media: User Perspectives on Moderation Choices, Interface Design, and Labor. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 289 (oct 2023), 33 pages. https://doi.org/10.1145/3610080

[28] Ridley Jones, Lucas Colusso, Katharina Reinecke, and Gary Hsieh. 2019. R/Science: Challenges and Opportunities in Online Science Communication. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300383

[29] Jukka Jouhki, Epp Lauk, Maija Penttinen, Niina Sormanen, and Turo Uskali. 2016. Facebook's emotional contagion experiment as a challenge to research ethics. *Media and Communication* 4, 4 (2016).

[30] Katherine Keith, David Jensen, and Brendan O'Connor. 2020. Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

[31] Emre Kiciman, Scott Counts, and Melissa Gasser. 2018. Using longitudinal social media analysis to understand the effects of early college alcohol use. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12.

[32] Charles Kiene and Benjamin Mako Hill. 2020. Who Uses Bots? A Statistical Analysis of Bot Usage in Moderation Teams. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3334480.3382960

[33] Sara Kiesler, Robert Kraut, and Paul Resnick. 2012. Regulating behavior in online communities. *Building Successful Online Communities: Evidence-Based Social Design* (2012).

[34] Gary King and Richard Nielsen. 2019. Why propensity scores should not be used for matching. *Political analysis* 27, 4 (2019), 435–454.

[35] Yubo Kou, Xinning Gui, Shaozeng Zhang, and Bonnie Nardi. 2017. Managing Disruptive Behavior through Non-Hierarchical Governance: Crowdsourcing in League of Legends and Weibo. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 62 (dec 2017), 17 pages. https://doi.org/10.1145/3134697

[36] Candice Lanius. 2019. Torment Porn or Feminist Witch Hunt: Apprehensions About the #MeToo Movement on /r/AskReddit. *Journal of Communication Inquiry* 43, 4 (2019), 415–436. https://doi.org/10.1177/0196859919865250 arXiv:https://doi.org/10.1177/0196859919865250

[37] Renkai Ma and Yubo Kou. 2021. "How Advertiser-Friendly is My Video?": YouTuber's Socioeconomic Interactions with Algorithmic Content Moderation. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 429 (oct 2021), 25 pages. https://doi.org/10.1145/3479573

[38] Renkai Ma and Yubo Kou. 2022. "I'm Not Sure What Difference is between Their Content and Mine, Other than the Person Itself": A Study of Fairness Perception of Content Moderation on YouTube. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 425 (nov 2022), 28 pages. https://doi.org/10.1145/3555150

[39] Renkai Ma and Yubo Kou. 2023. "Defaulting to Boilerplate Answers, They Didn't Engage in a Genuine Conversation": Dimensions of Transparency Design in Creator Moderation. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 44 (apr 2023), 26 pages. https://doi.org/10.1145/3579477

[40] Mike Masnick. 2019. Protocols, Not Platforms: A Technological Approach to Free Speech. https://knightcolumbia.org/content/protocols-not-platforms-a-technological-approach-to-free-speech

[41] J Nathan Matias. 2016. Posting rules in online discussions prevents problems & increases participation. *Civil Servant* 8 (2016).

[42] Nathan J. Matias. 2016. The Civic Labor of Online Moderators. In *Internet Politics and Policy conference* (Oxford, United Kingdom). Oxford, United Kingdom.

[43] Jim McCambridge, John Witton, and Diana R Elbourne. 2014. Systematic review of the Hawthorne effect: new concepts are needed to study research participation effects. *Journal of clinical epidemiology* 67, 3 (2014), 267–277.

[44] Aiden McGillicuddy, Jean-Gregoire Bernard, and Jocelyn Cranefield. 2016. Controlling Bad Behavior in Online Communities: An Examination of Moderation Work. *ICIS 2016 Proceedings* (dec 2016). http://aisel.aisnet.org/icis2016/SocialMedia/Presentations/23

[45] Jacob Metcalf and Kate Crawford. 2016. Where are human subjects in big data research? The emerging ethics divide. *Big Data & Society* 3, 1 (2016), 2053951716650211.

[46] Cornelia Moser. 2001. How open is' open as possible'?: three different approaches to transparency and openness in regulating access to EU documents. (2001).

[47] Alexandra Olteanu, Onur Varol, and Emre Kiciman. 2017. Distilling the Outcomes of Personal Experiences: A Propensity-Scored Analysis of Social Media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 370–386. https://doi.org/10.1145/2998181.2998353

[48] 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 19th International Conference on Supporting Group Work* (Sanibel Island, Florida, USA) *(GROUP '16)*. ACM, New York, NY, USA, 369–374. https://doi.org/10.1145/2957276.2957297

[49] Sarah T Roberts. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press.

[50] Donald B Rubin. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc.* 100, 469 (2005), 322–331.

[51] Adam Sadilek, Henry Kautz, and Vincent Silenzio. 2012. Modeling spread of disease from social interactions. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 6. 322–329.

[52] Koustuv Saha, Eshwar Chandrasekharan, and Munmun De Choudhury. 2019. Prevalence and psychological effects of hateful speech in online college communities. In *Proceedings of the 10th ACM Conference on Web Science*. 255–264.

[53] Koustuv Saha, Sindhu Kiranmai Ernala, Sarmistha Dutta, Eva Sharma, and Munmun De Choudhury. 2020. Understanding Moderation in Online Mental Health Communities. In *HCII*. Springer.

[54] Koustuv Saha, Yozen Liu, Nicholas Vincent, Farhan Asif Chowdhury, Leonardo Neves, Neil Shah, and Maarten W Bos. 2021. AdverTiming Matters: Examining User Ad Consumption for Effective Ad Allocations on Social Media. In *Proc. CHI*.

[55] Koustuv Saha and Amit Sharma. 2020. Causal Factors of Effective Psychosocial Outcomes in Online Mental Health Communities. In *ICWSM*.

[56] Koustuv Saha, Benjamin Sugar, John Torous, Bruno Abrahao, Emre Kıcıman, and Munmun De Choudhury. 2019. A social media study on the effects of psychiatric medication use. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 440–451.

[57] Koustuv Saha, Ingmar Weber, and Munmun De Choudhury. 2018. A Social Media Based Examination of the Effects of Counseling Recommendations After Student Deaths on College Campuses. In *Twelfth International AAAI Conference on Web and Social Media*.

[58] santaclaraprinciples. org. 2020. Santa Clara principles on transparency and accountability in content moderation. *Santa Clara Principles* (2020).

[59] Nathan Schneider, Primavera De Filippi, Seth Frey, Joshua Tan, and Amy Zhang. 2021. Modular Politics: Toward a Governance Layer for Online Communities. *Proc. ACM Hum.-Comput. Interact.* CSCW (Oct. 2021).

[60] Sarita Schoenebeck, Oliver L Haimson, and Lisa Nakamura. 2020. Drawing from justice theories to support targets of online harassment. *New Media & Society* (2020). https://doi.org/10.1177/1461444820913122

[61] Joseph Seering. 2020. Reconsidering Self-Moderation: The Role of Research in Supporting Community-Based Models for Online Content Moderation. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 107 (oct 2020), 28 pages. https://doi.org/10.1145/3415178

[62] 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* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 111–125. https://doi.org/10.1145/2998181.2998277

[63] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator engagement and community development in the age of algorithms. *New Media & Society* (2019), 1461444818821316.

[64] Manya Sleeper, Alessandro Acquisti, Lorrie Faith Cranor, Patrick Gage Kelley, Sean A. Munson, and Norman Sadeh. 2015. I Would Like To..., I Shouldn't..., I Wish I...: Exploring Behavior-Change Goals for Social Networking Sites. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Vancouver, BC, Canada) *(CSCW '15)*. Association for Computing Machinery, New York, NY, USA, 1058–1069. https://doi.org/10.1145/2675133.2675193

[65] Mark C. Stafford and Mark Warr. 1993. A Reconceptualization of General and Specific Deterrence. *Journal of Research in Crime and Delinquency* 30, 2 (1993), 123–135. https://doi.org/10.1177/0022427893030002001 arXiv:https://doi.org/10.1177/0022427893030002001

[66] Nicolas P Suzor, Sarah Myers West, Andrew Quodling, and Jillian York. 2019. What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation. *International Journal of Communication* 13 (2019), 18.

[67] Samuel Hardman Taylor, Dominic DiFranzo, Yoon Hyung Choi, Shruti Sannon, and Natalya N. Bazarova. 2019. Accountability and Empathy by Design: Encouraging Bystander Intervention to Cyberbullying on Social Media. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 118 (nov 2019), 26 pages. https://doi.org/10.1145/3359220

[68] Amaury Trujillo and Stefano Cresci. 2022. Make Reddit Great Again: Assessing Community Effects of Moderation Interventions on r/The_Donald. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 526 (nov 2022), 28 pages. https://doi.org/10.1145/3555639

[69] 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.

[70] Kristen Vaccaro, Christian Sandvig, and Karrie Karahalios. 2020. "At the End of the Day Facebook Does What It Wants": How Users Experience Contesting Algorithmic Content Moderation. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 167 (oct 2020), 22 pages. https://doi.org/10.1145/3415238

[71] Kristen Vaccaro, Ziang Xiao, Kevin Hamilton, and Karrie Karahalios. 2021. Contestability For Content Moderation. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 318 (oct 2021), 28 pages. https://doi.org/10.1145/3476059

[72] Lindy West. 2017. I've left Twitter. It is unusable for anyone but trolls, robots and dictators | Lindy West | Opinion | The Guardian. https://www.theguardian.com/commentisfree/2017/jan/03/ive-left-twitter-unusable-anyone-but-trolls-robots-dictators-lindy-west?CMP=share%7B%5C_%7Dbtn%7B%5C_%7Dtw

[73] Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society* (2018).

[74] Richard Ashby Wilson and Molly K Land. 2020. Hate speech on social media: Content moderation in context. *Conn. L. Rev.* 52 (2020), 1029.

[75] Sijia Xiao, Shagun Jhaver, and Niloufar Salehi. 2023. Addressing Interpersonal Harm in Online Gaming Communities: The Opportunities and Challenges for a Restorative Justice Approach. *ACM Trans. Comput.-Hum. Interact.* 30, 6, Article 83 (sep 2023), 36 pages. https://doi.org/10.1145/3603625

[76] Yunhao Yuan, Koustuv Saha, Barbara Keller, Erkki Tapio Isometsä, and Talayeh Aledavood. 2023. Mental Health Coping Stories on Social Media: A Causal-Inference Study of Papageno Effect. In *Proceedings of the ACM Web Conference 2023*. 2677–2685.

[77] Justine Zhang, Sendhil Mullainathan, and Cristian Danescu-Niculescu-Mizil. 2020. Quantifying the Causal Effects of Conversational Tendencies. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 131 (oct 2020), 24 pages. https://doi.org/10.1145/3415202