Content removal as a moderation strategy: Compliance and other outcomes in the ChangeMyView community

KUMAR BHARGAV SRINIVASAN, University of Colorado Boulder
CRISTIAN DANESCU-NICULESCU-MIZIL, Cornell University
LILLIAN LEE, Cornell University
CHENHAO TAN*, University of Colorado Boulder

Moderators of online communities often employ comment deletion as a tool. We ask here whether, beyond the positive effects of shielding a community from undesirable content, does comment removal actually cause the behavior of the comment's author to improve? We examine this question in a particularly well-moderated community, the ChangeMyView subreddit.

The standard analytic approach of interrupted time-series analysis unfortunately cannot answer this question of causality because it fails to distinguish the effect of *having made* a non-compliant comment from the effect of *being subjected to moderator removal* of that comment. We therefore leverage a "delayed feedback" approach based on the observation that some users may remain active between the time when they posted the non-compliant comment and the time when that comment is deleted. Applying this approach to such users, we reveal the causal role of comment deletion in reducing immediate noncompliance rates, although we do not find evidence of it having a causal role in inducing other behavior improvements. Our work thus empirically demonstrates both the promise and some potential limits of content removal as a positive moderation strategy, and points to future directions for identifying causal effects from observational data.

CCS Concepts: • **Applied computing → Law, social and behavioral sciences**; • **Human-centered computing → Collaborative and social computing**.

Additional Key Words and Phrases: content moderation, quasi-experimental designs, delayed feedback, interrupted time-series analysis, Reddit, ChangeMyView

ACM Reference Format:
Kumar Bhargav Srinivasan, Cristian Danescu-Niculescu-Mizil, Lillian Lee, and Chenhao Tan. 2019. Content removal as a moderation strategy: Compliance and other outcomes in the ChangeMyView community . *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 163 (November 2019), 21 pages. https://doi.org/10.1145/3359265

*Principal contact author.

We thank the anonymous reviewers and chairs for their very thoughtful feedback; their advice and questions resulted in substantial improvements to this paper. We also thank the members of the NLP+CSS research group at CU Boulder, Liye Fu, Bobby Kleinberg, and Jon Kleinberg for their insightful comments and discussions; and Kal Turnbull and the CMV moderators for providing access to moderation data. This material is based upon work supported by the National Science Foundation under grants BIGDATA SES-1741441 and CAREER IIS-1750615. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.
Authors' addresses: Kumar Bhargav Srinivasan, kumar.srinivasan@colorado.edu, University of Colorado Boulder, Boulder, Colorado, 80309; Cristian Danescu-Niculescu-Mizil, cristian@cs.cornell.edu, Cornell University, Ithaca, New York, 14850; Lillian Lee, llee@cs.cornell.edu, Cornell University, Ithaca, New York, 14850; Chenhao Tan, chenhao@chenhaot.com, University of Colorado Boulder, Boulder, Colorado, 80309.

2573-0142/2019/11-ART163 $15.00
https://doi.org/10.1145/3359265

1 INTRODUCTION

Moderators of online communities often remove inappropriate content in order to shield users from undesirable material and enforce community rules. But do authors of a deleted comment alter their behavior, for better or perhaps even for worse, after their comment is removed? In this work, we begin exploring this question by studying a particularly well-moderated and goal-oriented community, ChangeMyView.

The ChangeMyView (CMV) subreddit[1] hosts conversations where someone expresses a view, typically a controversial opinion, with the intention of hearing alternate perspectives; others then try to change that person's mind. Despite being fundamentally based on people arguing with each other, ChangeMyView has a reputation[2] for being remarkably civil and productive for an open internet site; here is what one journalist who joined it said:

> My opinion was no longer a "take" fitted to Twitter or an op-ed. It was a responsible perspective, honed in a collegial atmosphere. .... In a culture of brittle talking points that we guard with our lives, Change My View is a source of motion and surprise. Who knew that my most heartening ideological conversation in ages would involve gonads, gender wars, and for heaven's sake Reddit? [16]

Users consider moderator intervention to be one of the key factors behind the quality of discussion in CMV [16, 17], and comment removal is the moderators'[3] front-line strategy. (Removal is considered a public formal warning, which affected users can appeal; users are typically only banned from the site after 3 comment removals.) We aim to understand the effects of the considerable effort devoted by moderators to this activity (they manually removed 22,788 comments between January 2015 and March 2018) on the subsequent behavior of a deleted comment's author, or *affected individual*. There is significant subtlety, discussed in Section 3, in formulating concrete hypotheses that are testable with data currently available, that is, do not require changes to the moderators' modus operandi. Working under these constraints, we formulate a set of hypotheses taking the following template form:

> In CMV, for affected individuals that continue to participate in CMV, the removal *causes* a change for the better in X.

We adopt this "optimistic" template as a lens on whether moderators can view their work as not just clean-up (as important as that is), but also as a way to improve future behavior. Indeed, given that CMV is both a well-moderated community and one that stands out for its quality of discussions, it is sensible to investigate whether the latter feature is a consequence of the former. Our concrete hypotheses specify "change for the better in X" as follows:

H-NonCompliance: decrease in the rate of subsequent rule violations.

H-Toxicity: decrease in use of toxic language.

H-Achievement: increase in contributions achieving community goals (in the case of CMV, to persuade) and/or receiving community approval.

H-Engagement: increase in level of engagement.

To identify the causal effects of content moderation from observational data, we could start from a widely adopted method, interrupted time-series analysis (ITS) [14], with the removal of a comment serving as the "interruption". However, ITS does not allow us to discount the following counter-hypothesis: that the observed effects could be explained by the context surrounding the posting of a to-be-removed comment, rather than being caused by the moderation action itself.

[1] https://www.reddit.com/r/changemyview/. Recently, a spinoff website, https://changeaview.com, was created along roughly the same lines.

[2] Media coverage listed at https://changemyview.net/subreddit/#media-coverage.

[3] Non-moderators may downvote or report other people's content, but can't remove it.

Our innovation is to leverage a new quasi-experimental design that we call "delayed feedback"; this approach exploits the observation that *some affected individuals make comments in the (often hours-long) "pre-removal" window wherein the to-be-removed comment has not yet actually been removed.* With respect to such affected individuals,[4] we are able to identify the causal role of comment deletion in reducing immediate noncompliance rates, in comparison to "control" users as introduced in Section 3. This finding supports H-NonCompliance. For our three other hypotheses, although ITS reveals some significant changes in the behavioral aspect of interest before versus after comment removal, our delayed feedback analysis does not attribute these changes to the causal effect of the moderation action.

We conclude our work by discussing implications for content moderation and methodological implications for identifying causal effects from observational data.

2 BACKGROUND AND RELATED WORK

In this section, we first describe comment removal in CMV and review prior literature relevant to our four hypotheses. We then discuss enforcement and maintenance of social norms in general.

2.1 Comment Removal in CMV

CMV provides a detailed description of moderation standards and practices through a wiki page that is visible to all users[5]. A comment may be removed by a moderator if it is considered to violate one or more rules, such as by being rude or hostile, not challenging the original post, or violating any Reddit site-wide policy. Moderators in CMV tend to be very active in maintaining community standards, while at the same time not deleting comments based on topic or content as long as the comments follow the rules. Removed comments remain visible to moderators afterwards (unless the author removes it themselves). As an example, Figure 1 shows a screenshot of a publicly visible post tree[6] where moderator *ColdNotion* removed a comment from *D_DUB03*. The moderator usually replies to the removed comment and explains how to appeal.

Terminology. We use the following terminology throughout this paper:

- A **post tree** is a discussion or conversation tree, rooted at a top-level conversation-starting original post, and also containing the replies to the root post, replies to the replies, and so on as nodes.
- An **original poster (OP)** for a post tree is the author of its root.
- A **comment** is a non-root node in a post tree.
- **Affected** refers to the removal of a comment by a moderator, *where that removal is the first or second one experienced by the comment's author*, since the third removal within six months triggers a (temporary) ban. Note that the removal rate before the first removal is always 0%.
- An **affected tree** is a post tree within which a comment removal occurred.
- An **affected user** or an **affected individual** is the author of a removed comment.
- A **problematic** comment refers to a comment that would be deemed to violate community rules by moderators, either removed or not.

[4] It should be noted these users constitute a relatively high-activity subset of users with respect to the pre-removal window.

[5] https://www.reddit.com/r/changemyview/wiki/modstandards.

[6] https://www.reddit.com/r/changemyview/comments/9zx07n/cmv_it_should_be_legal_for_homeless_people_to/eadodfn/.

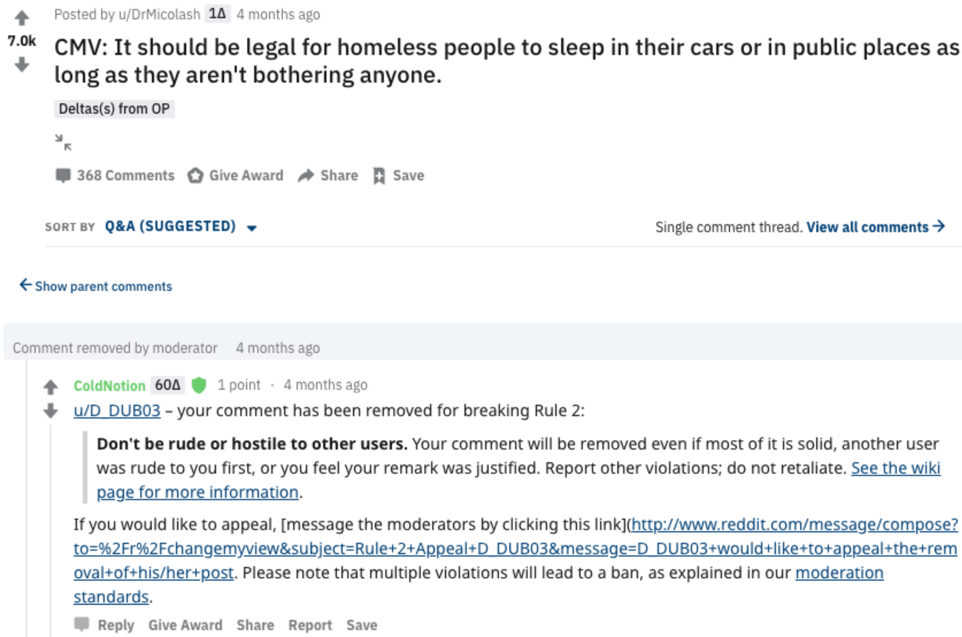


Fig. 1. Public display for a removed comment. A comment by user *D_DUB03* was removed by moderator *ColdNotion* for violating Rule 2 and replaced by the text "Comment removed by moderator". In this case, the moderator replied to the removed comment with the reason for removal and also included an explanation of how to appeal. Although the comment is no longer visible to the public, its content remains available in the moderator interface unless the user deletes it on their end.

2.2 H-NonCompliance: Related Work

Kiesler et al. [21] give a thorough consideration of the tangle of factors at play when it comes to whether or not regulating online behavior limits the effects or extent of bad behavior ("bad" is their choice of term). We enumerate here those of their design principles that relate to whether we can expect comment removal to reduce the rate of subsequent rule violations by the authors of removed comments, should they continue to participate in the CMV community. Kiesler et al.'s (henceforth KKRK) Design Claim 2 states that authors whose comments are deleted (rather than redirected to some other space) may post additional inflammatory content; on the other hand, CMV comment removals can be appealed, and thus, by KKRK Design Claim 3, will be more effective, since users react more positively to sanctions perceived to be administered fairly. KKRK Design Claim 7 advocates for a widely-adopted community norm of ignoring (rather than reacting to) trolls, which perhaps suggests that H-NonCompliance might not hold; moreover, the usual CMV comment-removal message violates KKRK Design Claim 23's recommendation to allow an affected individual to save face (e.g., by saying something like "you might not have known about this rule"). On the other hand, since CMV comment removal forms part of a sequence of more consequential moderation actions (that is, repeated removals lead to user bans), KKRK Design Claim 31's advocacy of graduated sanctions suggests support for H-NonCompliance. In all, we take these contrasting possibilities as evidence that there is no obvious a-priori answer as to whether H-NonCompliance holds, and thus it is an interesting hypothesis to explore.

When it comes to the offline world, the theory of specific deterrence — that a sanctioned individual will be deterred from the sanctioned behavior in the future[7] — has been shown to be insufficient, with many modulating factors at play [28, 34, 38], such as, as mentioned above, whether the affected individual perceives the process to be fair. Again, it does not appear possible to draw a direct prediction regarding H-NonCompliance from this literature.

Yet, inspired in part by (generalized) deterrence theory, Seering et al. [37] hypothesized that banning users that post spam messages on the Twitch gaming platform would reduce further spamming rates (by other users). While the spam rate actually increased, it increased by much less than during a comparable time period in which spammers were *not* banned. This observation contradicts H-NonCompliance when taken literally, but can be interpreted as supporting H-NonCompliance in spirit.

Our work on comment removal on CMV bears strong similarity to Chang and Danescu-Niculescu-Mizil [5]'s examination of the effect of temporarily banning users on Wikipedia, particularly with respect to the chance of future rule violations — it goes up, in contradiction of H-NonCompliance, but no causal link is claimed. Their focus differs from ours: they look at how an affected user's prior community engagement and perception of the fairness of the block (ban) are important mediating factors in post-block behavior.

2.3 H-Toxicity: Prior Work

Chancellor et al. [2] determine that after Instagram banned 17 tags used to highlight posts that advocate eating disorders,[8] variant tags arose among the pro-eating-disorder community that "depict more vulnerable, toxic, and 'triggering' content" (pg. 1209) — "toxic" meaning invoking self-loathing and self-harm — and the new variants were used more frequently than the old ones. Chancellor et al. [2] indeed recommend against suppressing content (and hence comment removal), due to negative consequences. Their work does not directly match H-Toxicity (tag ban vs. comment removal, collective behavior vs. behavior of the affected individual, Instagram vs. CMV, toxic=harmful vs. toxic=hate speech or swear words), but may still be taken as indirect evidence against H-Toxicity.

On the other hand, Chandrasekharan et al. [3] show that after Reddit banned several hate-based communities in 2015, participants in those communities that stayed active on Reddit collectively reduced their hate-speech usage by "at least 80%". Despite the differences with our setting (forum dissolution vs. comment removal, hate-group subreddits vs. CMV), we can interpret Chandrasekharan et al. [3]'s findings as support for H-Toxicity.

2.4 H-Achievement: Prior Work

With respect to achievement of community tasks, the first component of H-Achievement, we mention Cunha et al. [12]'s study of a weight-loss community that finds a causal correlation between number of support messages and a reported reduction in weight. We are unaware of work investigating the effects of negative feedback on accomplishment of specific goals. (Recall that CMV users strive to change someone's mind, operationally defined as receiving a "delta" badge on a comment.)

With respect to community approval, Cheng et al. [8] look at the subsequent reception of commenters' contributions after these commenters receive negative social feedback, finding that scores drop, indeed, even more than would seem warranted given computed text-quality scores

[7] *General* deterrence concerns whether the behavior of people other than the affected individuals improves.

[8] Banning tags is a softer form of moderation than comment removal: the content remains on the site, but searching on such tags bring up no results.

(which themselves also decrease). Similarly, Ahn et al. [1] show that the more downvotes a Stack Exchange user receives on a question, the lower the score of their subsequent questions. Although the settings of these two papers don't exactly match H-Achievement (online news/Q&A sites vs. CMV, low community-assigned scores vs. moderator deletion of comments), we interpret their results as contradicting H-Achievement with respect to its community-approval component.

While a less direct fit to hypothesis H-Achievement, the work of Cheng et al. [9] is also related. One of their experiments shows that having many comments deleted, as opposed to a few, leads to a drop in predicted community-approval score.

2.5 H-Engagement: Prior Work

Cheng et al. [8]'s social-feedback experiments, mentioned above, demonstrate increased commenting rates after receiving low community evaluations, an observation that supports H-Engagement. However, the temporary-ban study by Chang and Danescu-Niculescu-Mizil [5] on Wikipedia found no significant change in future activity rate.

2.6 Enforcement and Maintenance of Injunctive Norms: A Broader Landscape

The work we present in this paper focuses on a particular moderator reaction to the violation of explicitly stated community rules.[9] Community rules can be viewed as an instantiation of injunctive norms that characterize the perception of what most people approve or disapprove of (vs. descriptive norms that characterize the perception of what most people do) in the focus theory of normative conduct [11]. Prior research has studied the effect of *exposure* to (violating) such norms on community-level or discussion-level behavior [25–27]. For instance, through randomizing announcements of community rules to discussions in the Science subreddit, Matias [25] shows that making community rules visible prevents unruly and harassing conversations. In comparison, our focus is on the effect of moderator actions that *enforce* injunctive norms on individual users. It is worth remembering that the norms that moderators enforce can change based on circumstances, such as whether humorous content is allowed during a serious situation [23].

3 METHODS

Our hypotheses are explicitly phrased as causal, positing effects directly attributable to content moderation. However, we decline, for both practical and ethical reasons, to use randomized experiments to study these effects: we do not propose disrupting the functioning of a popular and productive community, nor do we advocate artificially or randomly altering the selection of comments to delete. Instead, in what follows, we describe quasi-experimental designs — both existing and new — that, although applied to observational data, can still provide partial insight into the causal effects of comment removal.

To orient our discussion, Figure 2A summarizes the set of possible situations we might ideally want to compare in order to address our causal hypotheses. These situations consist of possible states a user could be in around the time that they post a problematic comment. The top row corresponds to situations in which the user has posted the problematic comment, and either had this problematic comment removed by the moderators (Yes/Yes) or not (Yes/No). The bottom row corresponds to situations in which the user has not yet posted the problematic comment, and

[9] We mention a growing line of research on characterizing the rules themselves in online communities. Keegan and Fiesler [20] study the evolution of rules on Wikipedia by tracking revisions on rule-related Wikipedia pages. Fiesler et al. [13] provide a characterization of different types of rules and show that community rules share common characteristics across subreddits. In comparison, Chandrasekharan et al. [4] perform a large scale study to understand content moderation through the language used in the removed comments on Reddit and identify norms that are universal (macro), shared across certain groups (meso), and specific to individuals (micro).

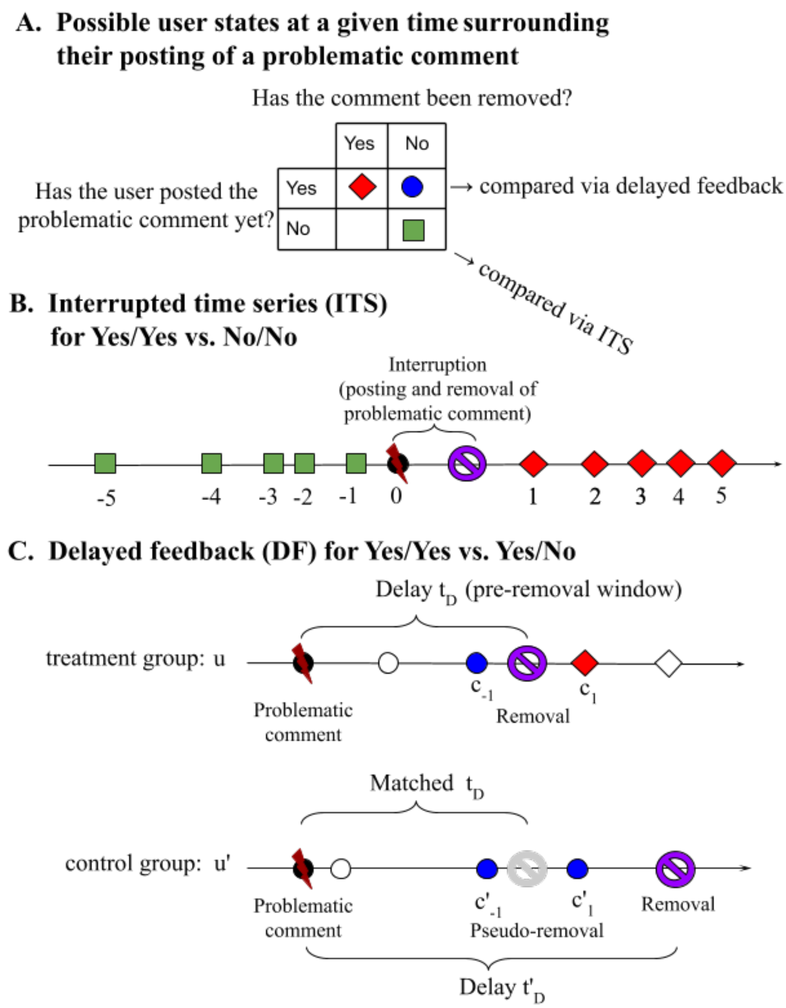


Fig. 2. **A.** User situations relevant to our causal hypotheses and how they are compared in our observational setting. **B.** Sketch of our use of interrupted time-series analysis (ITS) where the 'interruption' consists of the (joint) posting and removal of the problematic comment. **C.** Sketch of the delayed feedback (DF) design, where the 'delay' corresponds to the interval between the time of an action (posting of the problematic comment, in our case) and the time of 'feedback' on that action (its removal, in our case).

either had this problematic comment removed (No/Yes) or not (No/No). We do note that the No/Yes case cannot occur by definition, since moderators cannot remove a problematic comment that has not yet been made.[10] Moreover, in observational settings, some cases may not turn out to be directly comparable; for example, if the moderators never remove a certain type of problematic comment, then users who make those problematic comments will never be part of the data for the Yes/Yes quadrant. Nevertheless, we can find comparable and accessible sub-cases of the states

[10] We assume in this work that moderators do not delete non-problematic comments. It is beyond the scope of this paper, although potentially fruitful, to expand our consideration of possible states of interest to those where, say, a user has a non-problematic comment removed.

as categorized by Figure 2A that allow (a limited) investigation of our causal hypotheses through quasi-experimental designs.

We first (Section 3.1) discuss using the standard technique of interrupted time-series analysis, which, from the perspective of Figure 2A, compares situations occurring across the main diagonal (Yes/Yes vs. No/No). This exposes the joint effect on the user's behavior of writing the problematic comment and having the problematic comment removed. Second (Section 3.2), we aim to separate the effect of having the problematic comment removed from the effect of simply having made a problematic comment in the first place — that is, to see whether a user, after making the problematic comment, would behave the same whether or not their problematic comment is removed.[11] Towards this aim, we introduce a new *delayed feedback* approach; this methodology exploits the observation that there is a delay between the moment a problematic comment is posted and the moment it is removed by the moderators, allowing comparisons across Figure 2A's top row (Yes/Yes vs. Yes/No).

3.1 Interrupted Time-Series Analysis: A Standard Design

In order to observe the (joint) effect of writing a problematic comment and of its removal, while at the same time controlling for the characteristics of the user, we employ the standard quasi-experimental design of interrupted time-series analysis (ITS). As illustrated in Figure 2B, this design allows us to observe and compare between situations that fall across the diagonal of Figure 2A. A given user is in a Yes/Yes situation *after* the problematic comment they posted is removed (◆); the same user is in a No/No situation *before* they posted the problematic comment (■). We can then compare the behavior of the user in their first k comments after their problematic comment was removed with that of their last k comments before they posted the problematic comment ($k = 5$ in Figure 2B). A discontinuity in behavioral trends that occurs at the time the problematic comment is posted and removed can then be interpreted as being the result of this *interruption*.

Formally, for a given user u we can model the temporal variation of their observed behavior as

$$y(i) = \beta_0 + \beta_1 i + \beta_2 x(i) + \beta_3(i \cdot x(i)), \tag{1}$$

where $i \in \{-k, \ldots, -3, -2, -1, 1, 2, 3, \ldots, k\}$ indexes u's comments such that $i = -1$ corresponds to the user's last comment before the 'interrupting' problematic comment and $i = 1$ corresponds to the user's first comment after the 'interrupting' problematic comment.[12] The binary variable $x(i)$ indicates whether the observation is made pre-interruption ($x(i) = 0$ for $i < 0$) or post-interruption ($x(i) = 1$ for $i > 0$). The pre-interruption slope β_1 estimates the underlying trend in the absence of an interruption; the level change β_2 estimates the change in level that can be attributed to the interruption; and the change in slope β_3 quantifies the difference between the pre-interruption and post-interruption slopes [14, 22, 32]. By quantifying discontinuities in pre/post-interruption behavioral trends, this methodology accounts for temporal effects that might otherwise explain a change between the before and after measurements.

3.2 Delayed Feedback: A New Design

Despite being a broadly adopted quasi-experimental design, in our setting, standard ITS cannot show that the user, after making a problematic comment, would have behaved worse if the comment had not been removed — in which case, one cannot say that the removal *caused* behavior improvement. For example, a behavioral discontinuity could be explained by the context surrounding the posting

[11] For example, with respect to H-Toxicity, it could be that when a user posts a comment with strong swear words, they have usually reached maximum anger (or realize they've crossed a line), so their subsequent comments are less toxic regardless of whether they are officially warned.

[12] We ignore the problematic comment itself, and discard cases in which the user made comments between the problematic comment and the time of its removal, in order to avoid ambiguity with respect to the location of discontinuity.

of the problematic comment (e.g., the user's mood might have reached maximum anger when they posted the comment, so that subsequent comments are less toxic) rather than by the moderation action. In order to zero in on the effect of the removal itself, we need to compare across the top row of Figure 2A (Yes/Yes vs. Yes/No), i.e., to take situations where the user has posted a problematic comment, and compare states wherein the comment has been removed against states where it has not.

Since we can only know that a comment is problematic if it was removed by a moderator,[13] the main difficulty is to find users that posted a problematic comment that was not removed (i.e., users in the Yes/No scenario). Our main intuition, sketched in Figure 2C (top), is to exploit the fact that moderator actions are not immediate: in fact, in about 40% of the cases, it takes more than two hours for the removal to take place (Figure 3a). In particular, behavior observed after the user posted the problematic comment but before its removal — a delay period which we call the *pre-removal window*, denoting activity within it with a blue circle (●) — cannot be attributed to the effects of the moderation action that hasn't yet happened.

We could then estimate the effect of removal on the individual that posted the problematic comment by comparing their behavior after its removal (comment c_1, ◆) — a time when the user is in a Yes/Yes situation — with that exhibited in the pre-removal window (comment c_{-1},●), wherein even though the user had posted the problematic content, they had not yet experienced its removal (Yes/No). We need, however, to account for the possibility that the observed changes can be attributed to temporal effects (since c_{-1} always occurs before c_1, and users are expected to change behavior with time). A discontinuity analysis cannot be used to address this concern as in the ITS approach, due to the limited observation window between the posting and removal of the problematic comment.[14]

To account for temporal effects and check whether the change in behavior is indeed due to the moderation action, we rely on a temporally paired "control" group (Figure 2C bottom): for each user u that had the problematic comment removed with a delay of t_D after the posting time, we select a matched user u' that also had a comment removed, but with a slightly larger delay of $t'_D > t_D$.[15] We can then consider the difference in u''s behavior before and after the pseudo-removal, i.e., the difference between their behavior in c'_{-1} and in c'_1 — to indicate the underlying temporally-driven change that we can expect a user to undergo around t_D seconds after they made the problematic comment. By discounting this underlying change, we can isolate the effect of comment removal from that of temporal effects.[16]

Importantly, this methodology assumes the duration of the delay is not related to the behavior of interest; while in other domains this was shown to hold true [37], this assumption could be challenged in our domain. To mitigate this risk, we select u' such that their delay time t'_D is as close as possible to t_D. Furthermore, the validity of this control can be examined by comparing c'_{-1} and c_{-1}, where we would expect no difference in behavior between u' and u.

[13] We do not have the resources to manually check non-deleted comments for problematicity.

[14] On average, a user that has posted at least one comment in the pre-removal window contributed only 0.2 additional comments during that time window in non-affected trees (or 2.5 in affected trees).

[15] Borrowing terminology from randomized experiments into this quasi-experimental design, we will say that u' belongs to the "control" group and u belongs to the "treatment" group.

[16] Our choice of control for temporal effects is similar to that used by Oktay et al. [30]: they compare change in behavior before vs. after some "treatment" event with change in behavior before vs. after a matched pseudo-event. As an aside, it is interesting to note that their focus is the reverse of ours, in the sense that the "treatment" event is the action itself (in their case, the posting of an answer on a Q&A site) rather than the feedback this action receives (its selection as a high-quality answer); as such, they do not consider the action-to-feedback delay for matching treatment users with control users.

Number of post trees	73,047
Number of users in CMV who comment at least once	176,409
Number of comments in CMV	4,176,818
Number of comments removed	22,788
Number of moderators who removed comments	43
Number of users in CMV whose comment is removed at least once	12,481
Total post trees with removed comments	8,463

Table 1. Statistics of our dataset.

More general applicability. We note that this delayed-feedback quasi-experimental design can be applied more broadly to estimate the effect of other types of feedback, beyond the removal of a comment. By exploiting the *delay* between an action and the feedback that action receives, the approach can disentangle the effect of the feedback from that of the circumstances of the action itself.

4 APPLICATION TO THE CMV SETTING

We obtained all[17] 73K ChangeMyView post trees dating from the subreddit's inception, January 2013, until March 2018 from https://pushshift.io, a site maintained by Jason Baumgartner. CMV administrators gave us access to the content and meta-data for all 23K comments removed by the moderators during the same period for breaking CMV's rules.[18][19] The meta-data contains time of removal, URL, username, moderator name, rule violated and description of violation. Violations range from personal hostility to lack of substance (e.g., pure jokes, nothing but a URL). Table 1 provides basic statistics about the dataset. The public version of our dataset (excluding the content of the removed comments) is available at https://chenhaot.com/papers/content-removal.html, and will allow other researchers access to our filterings and matchings.

Our experiments consider only affected individuals that continue to participate in CMV after comment removal. It is important to point out that *affected individuals do abandon CMV at somewhat higher rates than other ChangeMyView members*, as shown in Figure 3b.[20] We cannot ascertain whether these community-departing users would have been productive or disruptive in CMV had they not incurred a removal. Affected individuals that do return to CMV typically do not return to the affected tree.

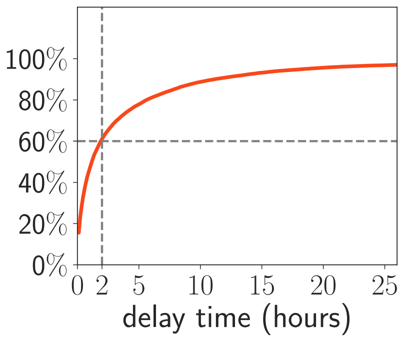
Applying the proposed approaches. We consider separately effects on the affected individuals' behavior in the affected tree and the behavior of those same individuals in non-affected trees, as the former is intrinsically tied to the context surrounding the making of the problematic comment, while the latter more broadly affects the community. Since interrupted time-series analysis requires extended observations before and after removal, as it happens, for data sparsity reasons we can only apply it to the non-affected tree scenario. We are able to apply the delayed feedback approach

[17] Throughout, references to "all" data are understood to be qualified by the possibility of inadvertent processing errors, Reddit API issues, and content deletion by users.

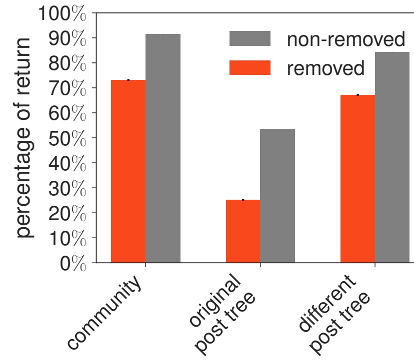
[18] Comment removal started in January, 2015, indicating opportunities for community-level natural experiments; pursuit of this idea lies beyond the scope of this paper.

[19] A reviewer asked whether CMV notifies users that their deleted comments are retained past 30 days. The CMV wiki entry on moderation standards (https://www.reddit.com/r/changemyview/wiki/modstandards) implies that the moderators have access to comment contents past 30 days, since otherwise they could not consider appeals of the "ban when 3 removals have occurred within 6 months" policy.

[20] This holds even when controlling for prior activity rate.



(a)



(b)

Fig. 3. (a) Cumulative distribution function of the delay between the time a problematic comment is posted and the time it is removed. (b) After making a comment c, to what CMV locations, if anywhere, does the author a post afterwards, depending on whether c was removed? Red: c was a's 1^{st} or 2^{nd} comment to ever be removed. Gray/stripes: all other comments. The same user can contribute to both colors via different comments. "community": the union of the original and different post trees. Error bars represent (tiny) standard errors.

in both non-affected trees and affected trees. Recall that interrupted time-series analysis allows us to study what happens after comment removal, while our delayed feedback design disentangles the effect of comment removal itself from other confounding factors; we compare user situations surrounding the posting of problematic comments associated with the first or second removal experienced by a user.

For ITS, we consider affected individuals who had 10 comments before and after removal in non-affected trees ($k = 10$), excluding the OP themselves and moderators. We do not include any users who made comments between posting the problematic comment and the time of its removal because the discontinuity point is ambiguous in that case. This procedure yields 2,752 instances of removals.

To apply the delayed feedback design to study behavioral changes in non-affected trees, we first identify every removal where the affected individual (u) made at least one comment during the pre-removal window and at least one comment within one week after removal (both comments are required to be in non-affected trees). For each of these instances, we try to find a matched control instance as detailed in Section 3.2, ensuring that the control affected individual (u') also has at least one comment before the matched pseudo-removal time, as well as at least one comment between this pseudo-removal time and the time of the actual removal of the problematic comment they posted. We choose matching instances for which the delay after comment time is as close as possible. Each removal can only be used as a control instance once. After discarding instances that cannot be matched, we are left with 775 treatment and control instances. We apply a similar procedure for the affected trees scenario by considering only comments that were posted in the affected trees, leading to 1,139 treatment and control instances.

Measuring behavioral changes. Table 2 summarizes all the features we used to operationalize our hypotheses. All features are measured based on the properties of a single comment. In our results figures in Section 5, the value depicted at a given comment index is the average feature value taken over all comments at that index.

Comment-level features	**Description**
H-NonCompliance: decrease in the rate of subsequent rule violations.	
Whether the comment is eventually removed	This directly measures whether a user violates community rules (as indicated by a moderator removing a comment).
H-Toxicity: decrease in use of toxic language.	
Swear-word ratio	Fraction of the comment's words that are swear words as defined by LIWC [33], capturing comment toxicity [7, 18]. Higher values indicate higher toxicity.
Hate-speech word ratio	Fraction of the comment's words that are hate-speech words as defined by the Hatespeech lexicon [15, 29].
H-Achievement: increase in contributions achieving community goals.	
Whether the comment wins a delta	A "delta" is a kind of positive feedback received by a user for changing the original poster's view. (We ignore deltas awarded by someone other than the OP.) Deltas indicate valuable contributions to the CMV community, which is dedicated to changing individuals' views.
Comment score	The difference between the number of upvotes and the number of downvotes. It represents a measurement of community feedback [39].
H-Engagement: increase in level of engagement	
Inter-comment time	The time gap between a comment and its next comment. This feature captures user engagement within the community: lower inter-comment time indicates higher engagement [19, 35].
Word count	Total number of words (excluding stopwords) in the comment, which serves as a proxy for the effort spent on writing a comment.
Depth	The depth of the comment as a node in the post tree, indicating how involved this user was in a conversation. Higher values indicate involved, detail-oriented discussions [10, 40, 41].

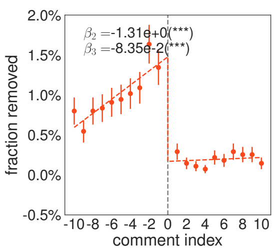
Table 2. Features considered in testing our hypotheses regarding user behavior changes after comment removal.

Ethical considerations. We acknowledge that, as discussed in Chandrasekharan et al. [4], the use of removed content from online communities for research purposes is controversial. Although we were granted moderator access to the content of removed comments, we will not share any information apart from what is publicly available on Reddit. (As Figure 1 shows, the content of the removed comment is not publicly available.) Although we have access to the username of the person making the removed comment (a username which is usually publicly displayed in the moderator's reply, as in Figure 1), our study does not focus on any single individual. We thus believe that the potential harm of our study is limited, and our findings will enable building better rules and governing principles for moderation within the CMV community.

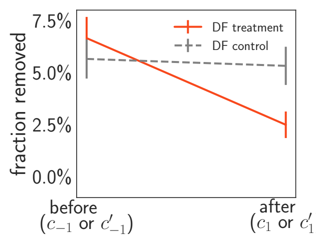
5 RESULTS

Next, we present results testing our four hypotheses formulated in the introduction.

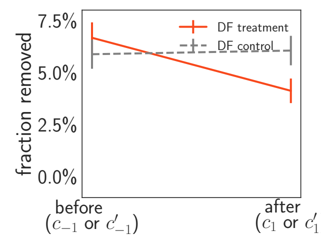
5.1 H-NonCompliance (Figure 4)



(a) Interrupted time-series analysis in non-affected trees (n=2,752).



(b) Delayed feedback in non-affected trees (↓↓↓, *, n=775).



(c) Delayed feedback in affected trees (↓↓, *, n=1,139).

Fig. 4. We use the following visualization scheme throughout the paper. In interrupted time-series analysis, $*s$ in the parentheses (if any) indicate p-values for significant changes via t-tests in the regression (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$). Our hypotheses can be supported by significant changes in either β_2 or β_3. In the delayed feedback approach, we use a t-test for evaluating whether user behavioral features in the first comment after removal (c_1) are significantly different from the last comment (c_{-1}) in the pre-removal window for the treatment group. Up-arrows indicate significance level of an *increase* in feature value; ↑↑↑: $p < 0.001$, ↑↑: $p < 0.01$; ↑: $p < 0.05$. Down-arrows show the same except for a *decrease* of feature value. We also use $*s$ to indicate whether the difference in difference ($(c_1 - c_{-1}) - (c'_1 - c'_{-1})$) is statistically significant (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$). Error bars indicate standard errors.

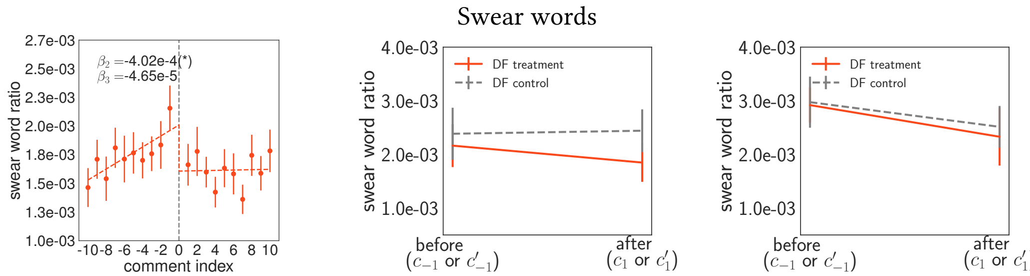
H-NonCompliance: ITS shows that noncompliance rate decreases after comment removal, and delayed feedback identifies the causal role of comment removal in reducing noncompliance rate both in non-affected trees and affected trees.

Our first hypothesis is concerned with the change in the rate of rule violations (*noncompliance rate*), measured by the fraction of comments removed. Ideally, moderators wish that noncompliance rate decreases after comment removal because comment removal is meant to serve as a warning to help members regulate their behavior rather than escalate the situation and lead to backfire effects.

Indeed, we observe a significant decrease in noncompliance rate after comment removal via interrupted time-series analysis, both for level and slope. In fact, the fraction of comments removed keeps going up before comment removal, which indicates that in addition to the comment removal performed by moderators, another plausible explanation for reduced noncompliance is due to reaching a local maximum in violating rules and affected individuals deciding to reset themselves.

Our delayed feedback approach allows us to zero in on the effect of comment removal itself. We find that noncompliance rate declines in the comment right after removal compared to the comment right before removal, both in non-affected trees and affected trees. In comparison, we do not observe a similar decrease in the control group. The fact that the treatment group and the control group have similar noncompliance rates before removal supports the validity of our matching scheme. Overall, these observations confirm the causal role of comment removal in reducing noncompliance rates, echoing Seering et al. [37] in spirit.

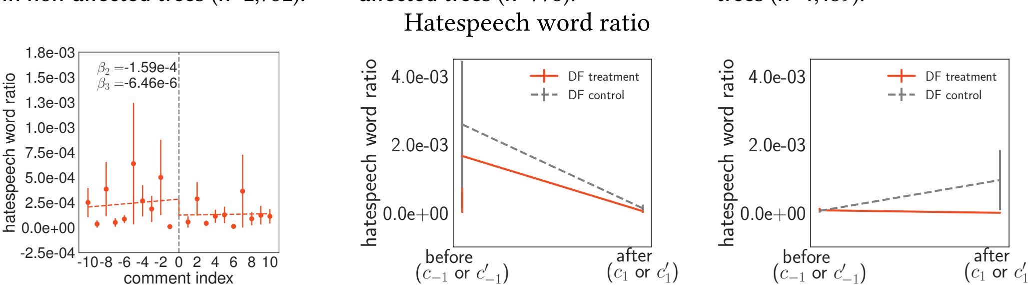
It is also important to note that the noncompliance rate in the delayed feedback approach is much higher than that in interrupted time-series analysis, which indicates that affected individuals that remained active in the pre-removal window may be generally more noncompliant than average affected individuals. Within the delayed feedback approach, the noncompliance rate is greater in affected trees than in non-affected trees, especially after comment removal. This suggests that



(a) Interrupted time-series analysis in non-affected trees (n=2,752).

(b) Delayed feedback in non-affected trees (n=775).

(c) Delayed feedback in affected trees (n=1,139).



(d) Interrupted time series in non-affected trees (n=2,752).

(e) Delayed feedback in non-affected trees (n=775).

(f) Delayed feedback in affected trees (n=1,139).

Fig. 5. **H-Toxicity:** Use of swear words significantly decreases after comment removal in ITS, but we do not observe any statistically change in the delayed feedback approach. None of the hate-speech changes are statistically significant, perhaps because the incidence rate is so low to begin with.

comment removal is more effective in regulating user behavior outside the situation that leads to the making of a removal-incurring comment.

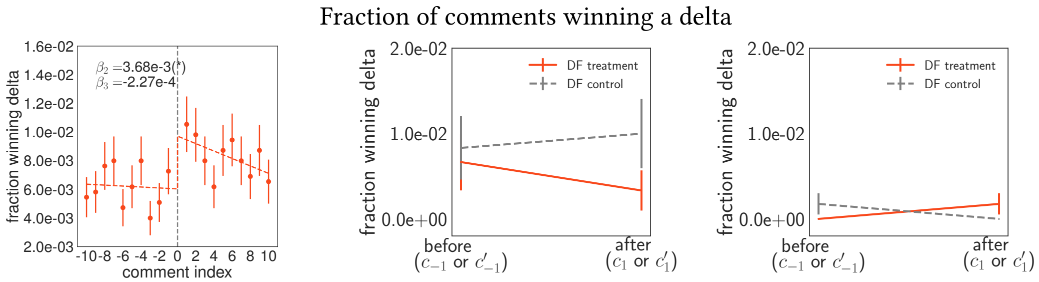
5.2 H-Toxicity (Figure 5)

Our second hypothesis is concerned with use of toxic language, another bad behavior that moderators may wish to discourage via comment removal. We focus on the use of swear words in this section, because hate speech turns out to be rarely used by affected individuals and none of the changes turn out to be statistically significant.

Interrupted time-series analysis shows a decrease in the use of swear words, the level change (β_2) (only) is statistically significant. However, this behavior change may not be caused by comment removal, because the delayed feedback approach shows no statistically significant change in non-affected trees or in affected trees, although there is a declining trend for the treatment group in both settings. Again, the fact that the treatment group and the control group are well matched in swear word ratio before removal supports the validity of our matching scheme. Overall, it seems that decreases in toxic-language use may not be induced by comment removal, at least for the comment immediately following comment removal.

5.3 H-Achievement (Figure 6)

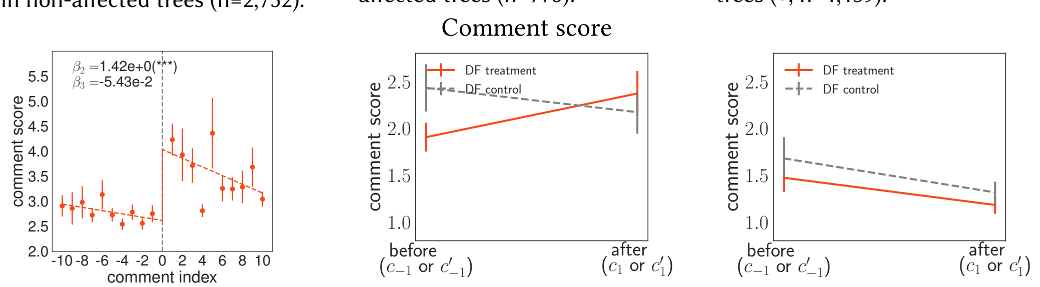
Our third hypothesis examines the good behavior that members can contribute to CMV, including making comments that actually change other users' views or that receive positive community feedback as measured by the difference between the number of upvotes and number of downvotes



(a) Interrupted time-series analysis in non-affected trees (n=2,752).

(b) Delayed feedback in non-affected trees (n=775).

(c) Delayed feedback in affected trees (∗, n=1,139).



(d) Interrupted time-series analysis in non-affected trees (n=2,752).

(e) Delayed feedback in non-affected trees (n=775).

(f) Delayed feedback in affected trees (n=1,139).

Fig. 6. **H-Achievement:** Although ITS shows significant increases in user achievement both in winning a delta and comment score, there is no significant change in any settings with the delayed feedback approach.

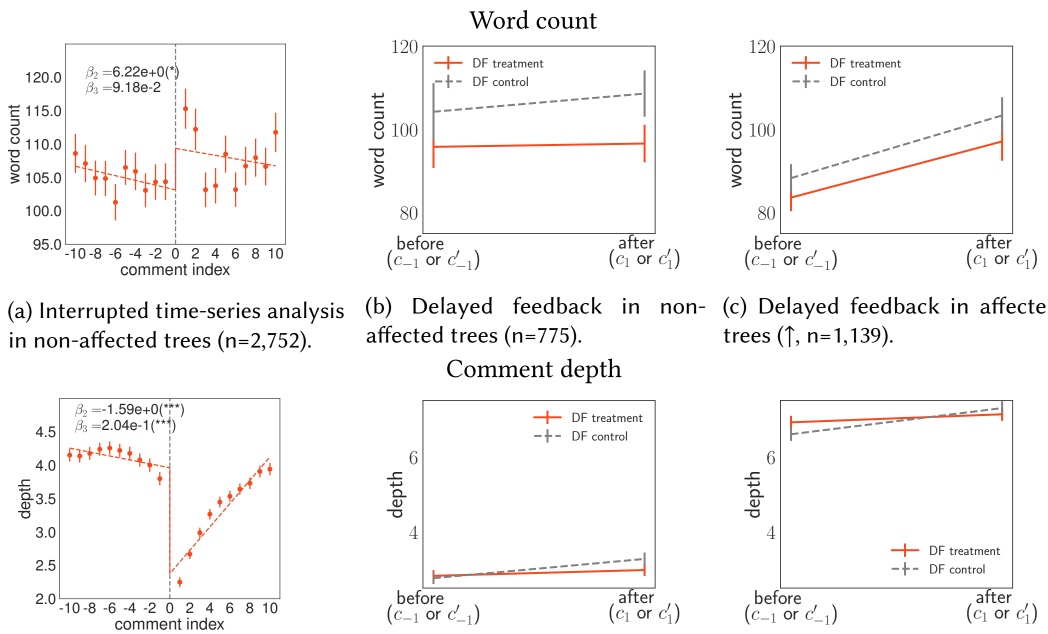
that a comment receives. Given that comment removal is a strategy for mitigating bad behavior, any positive effect on future good behavior would be notable for moderators.

Using interrupted time-series analysis, we observe a positive level change both in fraction of comments winning a delta and comment score, which suggests potential positive effects of comment removal. However, we cannot confirm the causal role of comment removal with our delayed feedback approach. Although the difference in difference is statistically significant in affected trees for winning a delta (Figure 6c), the treatment group and the control group are not well matched in the pre-removal window. The trends are also mixed for different combinations of achievement measures and non-affected/affected trees. This contrast suggests that the improvements in achievement might be due to temporal effects, such as getting out of the situation that leads to the making of a removal-incurring comment. Note that fraction of comments winning a delta in affected trees and comment score in non-affected trees are the only cases where the treatment group and the control group are not well matched in the pre-removal window.

5.4 H-Engagement (Figure 7 and Figure 8)

Our final hypothesis considers changes in level of engagement, such as posting more frequently and writing longer comments. Communities generally try to foster user engagement, although we caution that engagement should be considered more a side benefit than necessarily a goal in moderation on its own.

Using ITS, we find that word count increases after comment removal (statistically significant level change), and depth significantly changes after comment removal (a decline in the level and



(a) Interrupted time-series analysis in non-affected trees (n=2,752).

(b) Delayed feedback in non-affected trees (n=775).

(c) Delayed feedback in affected trees (↑, n=1,139).

(d) Interrupted time-series analysis in non-affected trees (n=2,752).

(e) Delayed feedback in non-affected trees (n=775).

(f) Delayed feedback in affected trees (**, n=1,139).

Fig. 7. **H-Engagement:** ITS shows a statistically significant level change in word count, and statistically significant level and slope changes for comment depth. In the delayed feedback approach, although the change in word count is statistically significant in the treatment group in affected trees, the control group also shows a significant change, which suggests that the change in the treatment group is likely due to a temporal effect.

an increase in the slope). This suggests that in non-affected trees, users move to discussions of shallow depth and then engage deeply again. In comparison, we do not observe significant changes in non-affected trees with the delayed feedback approach.

Note that while the change in word count for the treatment group is statistically significant in affected trees, the control group also shows a significant increase in word count, suggesting that this shift can be attributed to a temporal effect. For example, it could be the state of the discussion in which the treatment and control group users were participating, rather than comment removal, that leads to users making longer comments. This observation highlights the importance of controlling for temporal effects in the delayed feedback approach.

Although the change in depth for the treatment group is not statistically significant, the treatment group does not dive into deeper discussions in affected trees compared to the control group, in a statistically significant way.

Finally, using ITS, we find that comment frequency significantly increases after comment removal, both in the level and in the slope. While this observation is consistent with Cheng et al. [8], we cannot claim that this change is due to comment removal, because we cannot apply the delayed feedback approach with this metric (inter-comment time is ill-defined given only one comment).

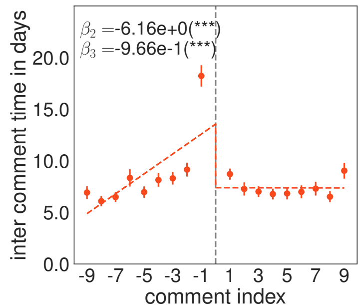


Fig. 8. **H-Engagement:** Interrupted time-series analysis shows that comment frequency significantly increases (inter-comment time decreases) after comment removal (n=2,752). Our delayed feedback approach cannot be applied to comment rate.

Hypothesis	ITS in non-affected trees: users with 10 comments before and after comment removal, and no comments in the pre-removal window, $n = 2,752$	DF in non-affected trees: users that posted in the pre-removal window and after removal in non-affected trees, and can be matched, $n = 755$	DF in affected trees: users that posted in the pre-removal window and after removal in affected trees, and can be matched, $n = 1,139$
H-NonCompliance	✓	✓	✓
H-Toxicity	✓		
H-Achievement	✓		
H-Engagement	✓		

Table 3. Summary of our key results.

6 CONCLUDING DISCUSSION

We have presented a systematic study of user behavior changes after comment removal, using the ChangeMyView community as a testbed. Our main contribution is to employ multiple methodologies to attempt to identify the causal role of the comment-removal moderation action. Specifically, in addition to the commonly-used interrupted time-series analysis, we propose a new delayed feedback quasi-experimental design to disentangle the effect of comment removal from that of the context surrounding the problematic comment being posted in the first place. Table 3 shows a summary of our key results. Interrupted time-series analysis shows that affected individuals not only regulate their bad behavior after comment removal, as reflected by reduced noncompliance and reduced toxicity in language, but also participate at increased rates and with increased achievement. Our delayed feedback approach, however, identifies a *causal* effect of comment removal only in reducing future noncompliance rates, not in inducing other behavioral improvements.

Implications for content moderation. Our results show that comment removal as a content moderation strategy is successful in regulating one type of bad behavior and potentially useful for encouraging good behavior in the CMV community, at least in the short term, since it consistently results in reduced noncompliance both when analyzed via interrupted time-series regression and when analyzed with our delayed feedback approach. Delayed feedback shows the results holding in both non-affected trees and affected trees, which suggests that comment removal does not (immediately) backfire even in the situation that leads to the making of a removal-incurring comment.

It is important to note that ITS and delayed feedback capture disjoint samples. The fact that noncompliance decreases for two disjoint samples further validates the effectiveness of comment removal in reducing rule violation in CMV.[21]

Furthermore, ITS reveals other positive behavior changes after removal, specifically, improvements in language non-toxicity, user achievement, and user engagement. However, since ITS does not account for the context of the problematic comment, we cannot causally attribute these effects to the moderation action.

Following the regulatory design discussions in Kiesler et al. [21], we see productive work ahead in exploring the following questions, and their corresponding design implications. Given our observation that noncompliance rate is much higher in affected trees than non-affected trees, should certain users be temporarily banned from contributing to the affected tree, rather than just warned via comment removal? For the affected individuals that are so discouraged by comment removal as to never return to the site, is there some way we could identify those who could potentially have been productive members if they had remained, and can we improve the return rate for such users by delaying the removal notification? Is it possible to predict diverse subsequent behavior of affected individuals and give them tailored moderation feedback? In our preliminary experiments, we find that it is feasible to forecast whether an affected individual will make another problematic comment based on user-level information, and linguistic cues have also been shown to signal future misbehavior [6, 24, 43]. Such predictive tools could, in the future, assist human moderators in their activities .

The latter two of the three previous questions also suggest a trade-off between the potential positive effect of enforcing community rules and the potential negative effect of driving people away with overzealous moderation. It is important to develop novel methods for understanding to what extent content moderation is necessary.

Finally, how does comment removal impact other users involved in the discussion? Although we have focused on users that are directly affected by comment removal in this work, it is important to understand the potential for spillover effects on the entire community. We are excited about the prospect of future research into these questions surrounding tools for algorithmically-assisted moderation. Meanwhile, we caution against automatic content moderation given the inherent biases and potentially complex consequences [31, 36, 42].

Methodological implications. Our study engages with methodological challenges in using observational data to understand the effects of content moderation. In the case of problematic-comment moderation, one should distinguish between the "interruption" comprised of the posting of the comment, the "interruption" comprised of its removal, and the "interruption" comprised of the passing of time. In order to disentangle the effect of comment removal from the circumstances surrounding the problematic comment's posting, we propose a delayed feedback approach that (i) takes advantage of the fact that users may remain active in the interval between their posting of a problematic comment and its removal by community moderators, and (ii) controls for the first and third "interruptions" just mentioned by finding users with a slightly longer delay between problematic-comment creation and removal.[22] It turns out that this control group is useful for filtering spurious results, e.g., a "finding" of an increase in word count after comment deletion.

[21]It is worth noting that in an early version of our delayed feedback approach, we observed an eventual increase in noncompliance rate when controls were not applied.

[22]Recall from Section 3.2 and the text surrounding footnote 14 that we cannot just apply ITS to the pre-removal vs. post-removal windows, due to data sparseness issues.

It is important to note that these methodological choices can affect characteristics of the samples that we end up analyzing. This sample selection effect can limit the generalizability of our findings. Moreover, sample-size considerations limit us, in the delayed feedback approach, to examining a single comment before and after removal, which means we can only capture affected individuals' immediate behavior changes.

Nevertheless, overall, we believe that delayed feedback represents a promising direction for future work in understanding the effect of content moderation, because a time gap always exists before manual content moderation occurs. As a community, we need to further develop observational methodologies for understanding important questions when ethical concerns prevent randomized experiments.

Limitations based on choice of community. The positive effect of comment removal could relate to the high-quality moderation of CMV and the commitment of the CMV community. In particular, the almost universally positive results with interrupted time-series analysis could relate to the fact that users with at least 20 comments in CMV are committed to the goals of CMV and are willing to reflect and make changes for the betterment of this community. It is also possible that comment removal in CMV applies a consistent standard (thus perceived as fair) that other communities may not emulate. Finally, CMV is a task-driven community with a clear goal of changing others' views and providing counterarguments. This goal may help moderators and members orient themselves, but other communities, such as news discussion forums, are not as goal-centered.

Examining the generalizability of our results to other online communities requires further work. If we observe different outcomes, it is useful to develop methods of characterizing the practice of moderation, such as how consistently moderators are able to apply the rules, and understand whether the type of moderation practice affects the outcomes.

Limitations based on data requirements of our methodology. As we discussed throughout the paper, selection bias is an important limitation of our experiments based on observational data. We could only run our analyses on users exhibiting sufficiently high levels of certain types of activity: we can make no claims regarding affected individuals that did not make "enough" comments in the targeted locations before and after the removal happened, including those that choose to leave the community after comment removal. Furthermore, given the anonymous nature of Reddit, it is possible that users switch to a different account after comment removal or being banned from the community; our approaches do not attempt to link accounts by the same user. It follows that our findings only apply to those users that remain active with the same account, and the positive behavior changes may not apply to people who choose to switch accounts. Another limitation is that due to the small sample size, we analyzed all comment removals in aggregate, without differentiating the reasons for removal.

Finally, we note that our study was not based on pre-established hypotheses, operationalization, and analysis plans, in the sense that our initial set of hypotheses have not heretofore been reported in this paper.[23] As such, our study should be considered exploratory. Nevertheless, our findings and delayed feedback methodology can inform the design of experiments that can further probe the causal nature of moderation actions.

[23] Our initial hypotheses were of the form, "comment removal causes a change in subsequent behavior", where 5 more characterizations of behavior (rates of hedging, type-to-token ratio, positive words, negative words, questions) were initially considered beyond the 8 reported in Table 2; following the reviewers' suggestions, we re-organized our hypotheses and in the process dropped those that did not fit well with this re-organization.

REFERENCES

[1] June Ahn, Brian S. Butler, Cindy Weng, and Sarah Webster. 2013. Learning to Be a Better Q'Er in Social Q&A Sites: Social Norms and Information Artifacts. In *Proceedings of the 76th ASIST Annual Meeting: Beyond the Cloud: Rethinking Information Boundaries (ASIST '13)*. American Society for Information Science, Silver Springs, MD, USA, Article 4, 10 pages. http://dl.acm.org/citation.cfm?id=2655780.2655784

[2] 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 CSCW (CSCW '16)*. ACM, New York, NY, USA, 1201–1213. https://doi.org/10.1145/2818048.2819963

[3] 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. *In Proceedings of ACM Human Computer Interaction* 1, Computer-Supported Cooperative Work and Social Computing, Article 31 (2017), 22 pages. http://doi.acm.org/10.1145/3134666

[4] 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. *In Proceedings of ACM Human Computer Interaction* 2, Computer-Supported Cooperative Work and Social Computing, Article 32 (Nov. 2018), 25 pages. http://doi.acm.org/10.1145/3274301

[5] Jonathan P. Chang and Cristian Danescu-Niculescu-Mizil. 2019. Trajectories of Blocked Community Members: Redemption, Recidivism and Departure. In *Proceedings of WWW*. 184–195. https://doi.org/10.1145/3308558.3313638

[6] Jonathan P. Chang and Cristian Danescu-Niculescu-Mizil. 2019. Trouble on the Horizon: Forecasting the Derailment of Online Conversations as they Develop. In *Proceedings of EMNLP*.

[7] Jilin Chen, Gary Hsieh, Jalal U. Mahmud, and Jeffrey Nichols. 2014. Understanding Individuals' Personal Values from Social Media Word Use. In *Proceedings of CSCW (CSCW'14)*. ACM, New York, NY, USA, 405–414. http://doi.acm.org/10.1145/2531602.2531608

[8] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2014. How community feedback shapes user behavior. In *Proceedings of ICWSM*. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8066

[9] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. Antisocial Behavior in Online Discussion Communities. In *Proceedings of International AAAI Conference on Web and Social Media(ICWSM)*. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10469/10489

[10] Daejin Choi, Jinyoung Han, Taejoong Chung, Yong-Yeol Ahn, Byung-Gon Chun, and Ted Taekyoung Kwon. 2015. Characterizing Conversation Patterns in Reddit: From the Perspectives of Content Properties and User Participation Behaviors. In *Proceedings of the 2015 ACM on Conference on Online Social Networks (COSN '15)*. ACM, New York, NY, USA, 233–243. http://doi.acm.org/10.1145/2817946.2817959

[11] Robert B. Cialdini, Carl A. Kallgren, and Raymond R Reno. [n. d.]. A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior.

[12] Tiago Cunha, Ingmar Weber, and Gisele Pappa. 2017. A Warm Welcome Matters!: The Link Between Social Feedback and Weight Loss in /R/Loseit. In *Proceedings of WWW Companion*. 1063–1072. https://doi.org/10.1145/3041021.3055131

[13] Casey Fiesler, Jialun Jiang, Joshua McCann, Kyle Frye, and Jed Brubaker. 2018. Reddit rules! Characterizing an Ecosystem of Governance. In *Proceedings of International AAAI Conference on Web and Social Media(ICWSM)*. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17898

[14] Antonio Gasparrini, Steven Cummins, and James Lopez Bernal. 2016. Interrupted time series regression for the evaluation of public health interventions: a tutorial. *International Journal of Epidemiology* 46, 1 (06 2016), 348–355. https://doi.org/10.1093/ije/dyw098

[15] Hatebase. 2018. Hatebase. https://www.hatebase.org/. [Online; accessed 3-Nov-2018].

[16] Virgina Heffernan. 2018. Our best hope for civil discourse online is on ... Reddit. *Wired* (2018). https://www.wired.com/story/free-speech-issue-reddit-change-my-view/

[17] Shagun Jhaver, Pranil Vora, and Amy Bruckman. 2017. *Designing for civil conversations: lessons learned from ChangeMyView.* Technical Report. Georgia Institute of Technology. https://smartech.gatech.edu/handle/1853/59080

[18] Shan Jiang and Christo Wilson. 2018. Linguistic Signals Under Misinformation and Fact-Checking: Evidence from User Comments on Social Media. *Proceedings of the ACM on Human-Computer Interaction* 2, Computer-Supported Cooperative Work and Social Computing, Article 82 (Nov. 2018), 23 pages. http://doi.acm.org/10.1145/3274351

[19] Andreas Kaltenbrunner, Vicenç Gómez, Ayman Moghnieh, Rodrigo Meza, Josep Blat, and Vicente López. 2007. Homogeneous temporal activity patterns in a large online communication space. *CoRR* abs/0708.1579 (2007). http://arxiv.org/abs/0708.1579

[20] Brian Keegan and Casey Fiesler. 2017. The Evolution and Consequences of Peer Producing Wikipedia's Rules. In *Proceedings of ICWSM*. https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15698

[21] Sara Kiesler, Robert E. Kraut, Paul Resnick, and Aniket Kittur. 2011. Regulating behavior in online communities. MIT Press, Cambridge, 125–178. http://www.jstor.org/stable/j.ctt5hhgvw.7

[22] Evangelos Kontopantelis, Tim Doran, David A Springate, Iain Buchan, and David Reeves. 2015. Regression based quasi-experimental approach when randomisation is not an option: interrupted time series analysis. *BMJ* 350 (2015). https://www.bmj.com/content/350/bmj.h2750

[23] Alex Leavitt and Joshua A. Clark. 2014. Upvoting hurricane Sandy: event-based news production processes on a social news site. In *Proceedings of CHI*. ACM, 1495–1504.

[24] Ping Liu, Joshua Guberman, Libby Hemphill, and Aron Culotta. 2018. Forecasting the presence and intensity of hostility on Instagram using linguistic and social features. In *Proceedings of ICWSM*.

[25] J. Nathan Matias. 2019. Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences* 116, 20 (2019), 9785–9789.

[26] Jonathan T. Morgan and Anna Filippova. 2018. 'Welcome' Changes?: Descriptive and Injunctive Norms in a Wikipedia Sub-Community. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018). https://doi.org/10.1145/3274321

[27] Lev Muchnik, Sinan Aral, and Sean J Taylor. 2013. Social influence bias: A randomized experiment. *Science* 341, 6146 (2013), 647–651. http://science.sciencemag.org/content/341/6146/647

[28] Daniel S. Nagin, Francis T. Cullen, and Cheryl Lero Jonson (Eds.). 2018. *Deterrence, choice, and crime: Contemporary perspectives*. Routledge, an imprint of the Taylor & Francis Group, New York, NY.

[29] Rishab Nithyanand, Brian Schaffner, and Phillipa Gill. 2017. Measuring Offensive Speech in Online Political Discourse. *CoRR* abs/1706.01875 (2017). http://arxiv.org/abs/1706.01875

[30] Hüseyin Oktay, Brian J. Taylor, and David D. Jensen. 2010. Causal discovery in social media using quasi-experimental designs. In *Proceedings of the First Workshop on Social Media Analytics*. ACM, 1–9.

[31] Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing Gender Bias in Abusive Language Detection. In *Proceedings of EMNLP*. https://www.aclweb.org/anthology/D18-1302

[32] Umashanthi Pavalanathan, Xiaochuang Han, and Jacob Eisenstein. 2018. Mind Your POV: Convergence of Articles and Editors Towards Wikipedia's Neutrality Norm. *In Proceedings of ACM Human Computer Interaction* 2, Computer-Supported Cooperative Work and Social Computing (2018). http://doi.acm.org/10.1145/3274406

[33] James W. Pennebaker, Roger J. Booth, and Martha E. Francis. 2007. Linguistic Inquiry and Word Count (LIWC): A computerized text analysis program. http://www.liwc.net/.

[34] Travis C. Pratt, Francis T. Cullen, Kristie R. Blevins, Leah E. Daigle, and Tamara D. Madensen. 2006. The Empirical Status of Deterrence Theory: A Meta-Analysis. In *Taking stock: The status of criminological theory*. Transaction Publishers, Piscataway, NJ, US, 367–395.

[35] Mattia Samory and Tanushree Mitra. 2018. Conspiracies Online: User Discussions in a Conspiracy Community Following Dramatic Events. In *Proceedings of ICWSM*. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17907

[36] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of ACL*.

[37] Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. Shaping pro and anti-social behavior on Twitch through moderation and example-setting. In *Proceedings of CSCW*. ACM, New York, NY, USA. http://doi.acm.org/10.1145/2998181.2998277

[38] Lawrence W. Sherman. 1993. Defiance, deterrence, and irrelevance: A theory of the criminal sanction. *Journal of Research in Crime and Delinquency* 30, 4 (1993), 445–473. https://doi.org/10.1177/0022427893030004006 01245.

[39] Chenhao Tan and Lillian Lee. 2015. All Who Wander: On the Prevalence and Characteristics of Multi-community Engagement. In *Proceedings of WWW*. https://doi.org/10.1145/2736277.2741661

[40] Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of WWW*. https://doi.org/10.1145/2872427.2883081

[41] Tim Weninger, Xihao Avi Zhu, and Jiawei Han. 2013. An Exploration of Discussion Threads in Social News Sites: A Case Study of the Reddit Community. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '13)*. ACM, New York, NY, USA, 579–583.

[42] Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. In *Proceedings of NAACL*.

[43] Justine Zhang, Jonathan P. Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Nithum Thain, and Dario Taraborelli. 2018. Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of ACL*. https://www.aclweb.org/anthology/P18-1125