

Algorithmic Pre-Moderation Reduces Toxicity by Changing What Users Write: Evidence from a Randomized Experiment in Newspaper Comments

Laura Bronner¹, Francisco Tomas-Valiente¹, Nicolai Berk¹, Karsten Donnay², Gloria Gennaro³, Fabrizio Gilardi², Philip Grech¹, Dominik Hangartner^{1*}

September 21, 2026

Abstract

Increasingly, online platforms have shifted from pre-publication content moderation—screening posts before they appear—to post-publication review that relies on user reports and subsequent moderator removal. Yet, rigorous evidence on how pre- and post-moderation affect discourse quality and participation remains limited. We report results from a randomized field experiment conducted with a leading Austrian newspaper featuring highly active user comment sections and operating under a default post-publication moderation regime. Articles were randomly assigned either to this post-moderation condition or a treatment that added algorithmic pre-moderation, screening comments before publication while leaving post-publication review unchanged. Across 410 articles and 115,657 submitted comments, assignment to pre-moderation reduced the share of toxic comments among ultimately published content from 4.5% to 3.4%, a 23% decline (p -value = 0.005). Importantly, this reduction is driven primarily by a decline in toxicity among submitted comments, indicating that users change what they write, not merely which comments are removed before publication. Engagement—the number of comments and unique commenters per article—remained unchanged. Comparing the same users across articles shows that individuals submit less toxic comments under pre-moderation than under post-moderation alone, consistent with users matching their contributions to the more civil tone of pre-moderated discussions. As platforms increasingly rely on post-moderation, our findings demonstrate the benefits of algorithmic pre-moderation and provide guidance for platforms and policymakers seeking to improve public discourse by preventing toxic content from setting the tone.

¹ Public Policy Group, ETH Zurich, Switzerland

² University of Zurich, Switzerland

³ University College London, United Kingdom

* Corresponding author. Email: dominik.hangartner@gess.ethz.ch

Keywords. Content moderation, public discourse, randomized field experiment, online toxicity, platform governance

Significance statement. How online platforms moderate user content has become a central question for democracy, free expression, and digital governance. Yet there is little causal evidence on how different moderation regimes shape online discourse. In a large-scale randomized field experiment with a national newspaper, we show that algorithmically screening comments before publication substantially reduces toxic content without reducing participation. Importantly, the effect arises because users adjust what they write in response to a more civil discussion tone, rather than because more content is mechanically filtered out before becoming visible. These findings enhance our understanding of the behavioral mechanisms underlying content moderation and provide actionable guidance for the design and governance of digital public spaces.

1 Introduction

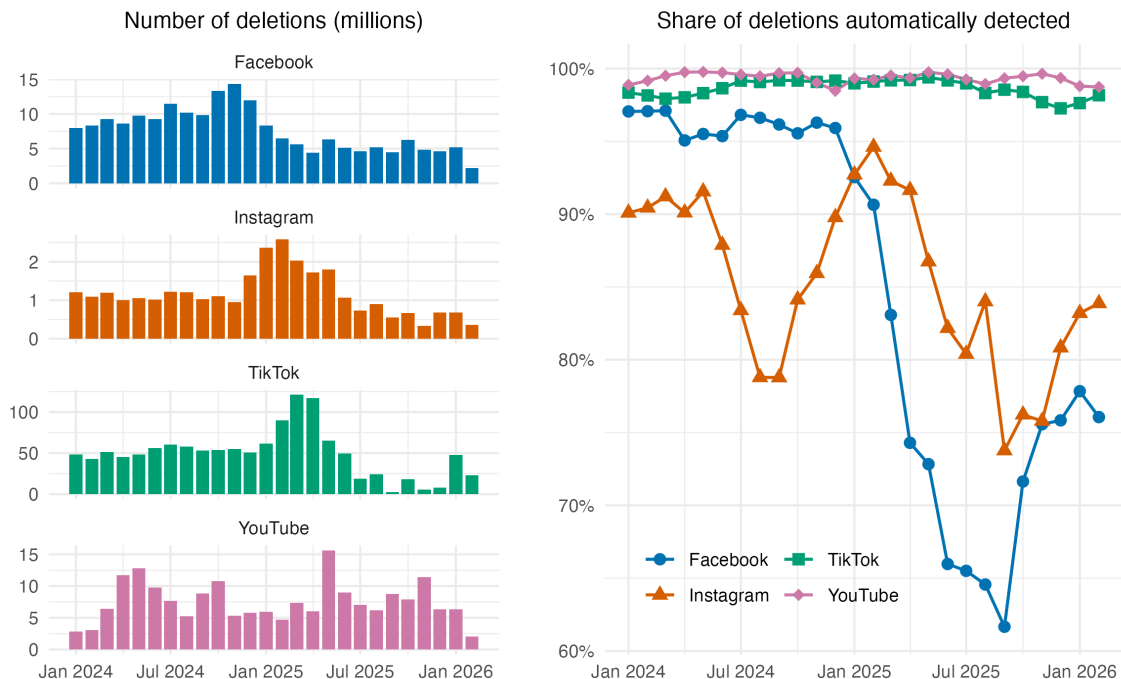
Harmful online speech degrades public discourse, distorts access to information, discourages participation, and can spill over into offline harm, raising urgent questions about how platforms should moderate comments [1–8]. Content moderation policies therefore play a central role in shaping the environment people encounter, both digitally and beyond. Online platforms used to rely heavily on *pre-moderation*—a screening process, often partially or wholly algorithmic, that preemptively deletes, demotes or holds comments before publication. Its potential benefits are wide-ranging: mechanically, it can prevent harmful content from ever being published, protecting users from the start; this mechanical effect can then induce users to behave differently—i.e. submit better comments—if they match the prevailing tone they encounter [9, 10], improving the conversation from the start and reducing the overall moderation load. At the same time, pre-moderation is criticized for cost, delay, and the risks of over-moderation and political bias [11–15]. For these reasons, platforms like X (formerly Twitter) and Meta have recently reduced their reliance on pre-moderation and instead turned more to *post-moderation*, allowing most content to be published and then relying on user reports or moderator review to remove harmful content after publication [e.g., 16–19].

Figure 1 documents this shift using data from the European Digital Services Act (DSA) transparency database. The left panel shows that, for three (Facebook, Instagram, TikTok) of the four major platforms, the total number of content deletions declined substantially since reporting began in 2024.¹ Consistent with a move away from pre-publication screening, the right panel shows that this decline in deletions coincided with a marked reduction in the share of removals initiated by automated systems—tools typically used for pre-moderation—particularly on Facebook and Instagram. As Meta announced in January 2025: “Up until now, we have been using automated systems to scan for all policy violations, but this has resulted in too many mistakes and too much content being censored that shouldn’t have

¹This decline is not driven by reduced engagement, which, if anything, increased during the same period [20]. Twitter/X is excluded because the shift to post-moderation occurred prior to 2024, and the DSA reporting data appear incomplete.

been. [...] For less severe policy violations, we’re going to rely on someone reporting an issue before we take any action” [21]. Despite this large-scale shift in moderation practices— affecting millions of users—experimental evidence on how pre- and post-moderation regimes shape public discourse is limited.

Figure 1: Reported content deletions and automated detections on major platforms



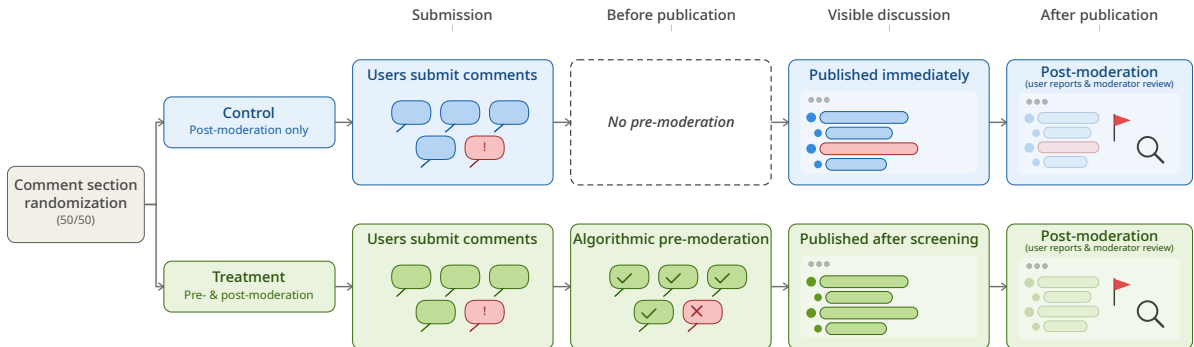
Note: The left panel plots the total number of monthly content deletions reported to the DSA transparency database, by platform. The right panel plots the share of all reported deletions that was detected by automated detection systems.

Existing research largely describes moderation practices [22, 23], uses observational designs to examine algorithmic pre-moderation [24], or evaluates *additional* ex-post moderation layered onto platforms’ existing systems [e.g., 11, 25, 26]. Complementary research studies how users experience moderation [27–29] and which approaches they prefer [14, 30–34]. While this literature has yielded important insights, there remains a lack of experimental evidence on two fundamental questions: First, does algorithmic pre-moderation improve discussion quality beyond mechanically filtering harmful content—by changing what users choose to write? Second, does reactive post-moderation improve the quality of online debate as much

as pre-moderation? Answering these questions is essential not only for advancing our theoretical understanding of content moderation mechanisms, but also for informing platform design and governance.

This scarcity of evidence reflects the difficulty of running experiments in naturalistic settings: To evaluate the effects of moderation regimes on user-produced content, researchers need to observe the full universe of posts—including those deleted or never published—to study mechanisms as well as outcomes. With some noteworthy exceptions [e.g. 35–38], large platforms rarely allow external research access of this kind. We overcome this challenge by partnering with *Der Standard*, Austria’s third-largest online news outlet [39], which hosts active comment sections below most articles with more than 50,000 submitted comments per day on average. We co-designed a field experiment to randomly vary a prompt for the newspaper’s moderation team to rely solely on post-moderation (control) or deploy algorithmic pre-moderation in addition to post-moderation (treatment) within a given article’s comment section (see Figure 2). For research-ethics and practical reasons, we implemented an encouragement design, leaving the ultimate decision on each article’s moderation regime to the human moderators, to allow for deviation from the assigned setting in the case of sensitive topics. Our intention-to-treat analysis accounts for this source of noncompliance with the experimental protocol.

Figure 2: Workflow for control and treated articles



Note: Comments can also be accepted, manually screened or rejected before publication if they are made by users who are new or on a black, grey or white list. See SI Section 2.4 for details.

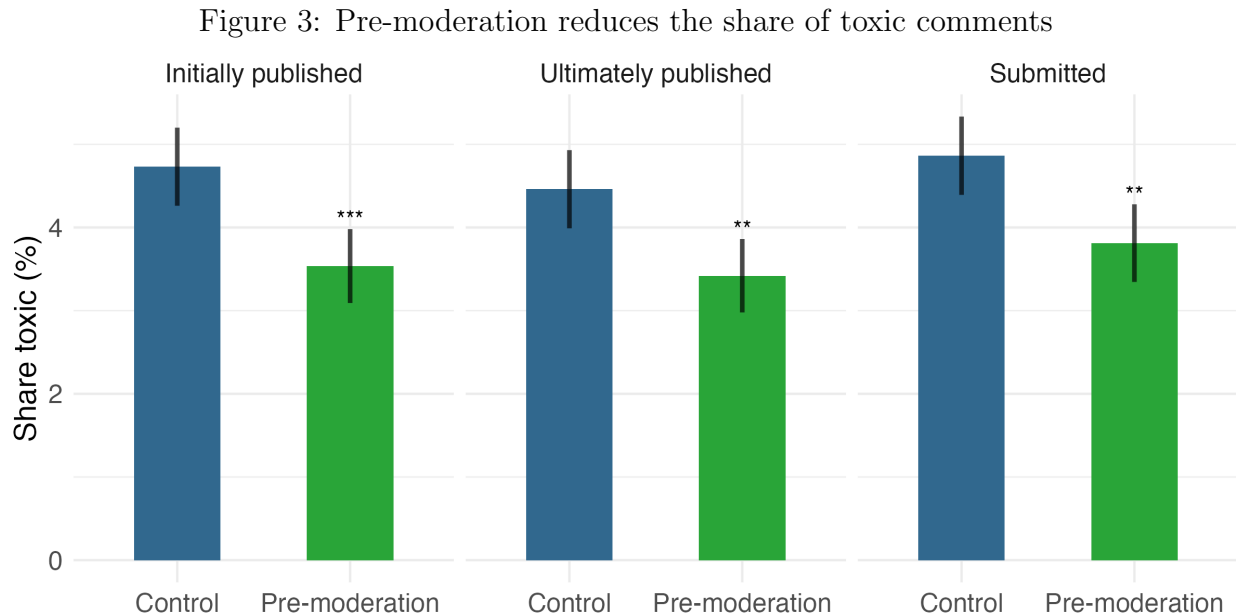
During the field phase (April 2024–March 2025), we randomly assigned 869 articles with a total of 320,798 submitted comments. To minimize frictions, moderators retained discretion over when, after publication, to randomize an article. Because some articles were randomized with a delay of several hours, our main estimates focus on comments submitted after randomization. As expected, articles in the treatment and control conditions are balanced with respect to the time passed between article publication and randomization. To maximize power, our main sample focuses on 410 articles (with a total of 115,657 comments) that were randomized within 95 minutes of publication, with robustness to alternative inclusion cut-offs and complier-based estimates on the full article sample (SI Section 6.1). We also exploit variation in the timing of randomization to illuminate mechanisms. As main outcomes, we report the share of toxic comments among (i) *initially* published comments (before any removals due to post-moderation), (ii) *ultimately* published comments (after post-moderation, measured five days after article publication), and (iii) all *submitted* comments, to separate exposure from behavior. Toxicity is measured with a BERT classifier fine-tuned on two datasets of German-language newspaper comments [see 40]. We also track engagement as the number of comments and unique commenters per article. Materials and Methods (MM) below provide further details on the sample, experimental design, outcomes, and statistical analysis. All statistical analyses, unless labeled as “exploratory”, were preregistered on OSF before we obtained the data.

2 Results

2.1 Main Results

Focusing on our main analysis sample, we first estimate the effects of the pre-moderation prompt on the toxicity of initially published comments, ultimately published comments, and all submitted comments. The left panel of Figure 3 shows that pre-moderation reduces the share of toxic comments among *initially* published comments from 4.7% to 3.5%, relative

to articles relying only on post-moderation (unadjusted p-value of the estimated difference in winsorized group means is $p < 0.001$; Benjamini-Hochberg-corrected $p_{BH} = 0.001$. See MM for details on statistical methods and Benjamini-Hochberg multiple-testing corrections). This 1.2-percentage-point reduction corresponds to a 25% decline in the share of published toxic comments—a substantial improvement in visible discourse quality.



Note: +, *, **, *** denote Benjamini-Hochberg-corrected statistical significance of the differences in means at the 10%, 5%, 1% and 0.1% levels, respectively. Outcomes were winsorized at the 2nd and 98th percentiles.

These gains largely persist after post-moderation, measured five days after article publication. The middle panel of Figure 3 shows that, among *ultimately* published comments, pre-moderation reduces the share of toxic comments from 4.5% in control articles to 3.4% ($p = 0.002$, $p_{BH} = 0.005$), a 23% decline. Consistent with this, pre-moderation also reduces the share of toxic content among comments that users visibly interact with, from 4.9% to 3.7% ($p = 0.001$, $p_{BH} = 0.008$; exploratory analysis, see SI 7.2 for details).

As registered in our pre-analysis plan, we expected post-moderation to substantially narrow treatment–control differences. Instead, the effect size remains fairly constant: the share of toxic comments among ultimately published comments in pre-moderated articles is 1.0

percentage point lower than in the control group, while the gap is 1.2 percentage points among initially published comments. We can understand this muted effect by comparing the control arms across the first two panels of Figure 3, which shows that the toxic share falls only slightly after post-moderation, from 4.7% among initially to 4.5% among ultimately published comments. In treated articles, the decrease is similarly modest (3.5% among initially to 3.4% among ultimately published comments). In other words, most toxic comments that are initially published remain online, and post-moderation does little to offset harmful exposure.

Additional evidence in SI Figure 28 further helps explain why post-moderation fails to compensate for the absence of pre-moderation. First, the share of initially published comments that users report is very low and statistically similar across moderation regimes (1.3% with pre-moderation versus 1.7% without; $p = 0.10$, $p_{BH} = 0.21$). Even though toxic comments are reported at higher rates than non-toxic ones (5.1% versus 1.5%), reporting remains sufficiently rare that most published toxic comments are never flagged by any user. Second, the share of initially published comments that are later deleted is almost identical across arms (3.5% with pre-moderation versus 3.8% without; $p = 0.65$, $p_{BH} = 0.70$). Taken together, these results indicate that, in our experiment, post-moderation cannot substitute for algorithmic screening largely because user reporting fails to surface most toxic content.

If post-moderation does not meaningfully reduce exposure to toxic content, how does pre-moderation work? Crucially, we find that the reduction in toxicity is not driven solely or even primarily by the mechanical removal of harmful content, but rather reflects changes in user behavior. The right panel of Figure 3 shows that pre-moderation reduces the share of toxic comments among *submitted* comments, from 4.9% to 3.8% ($p = 0.002$, $p_{BH} = 0.005$). This 1.1-percentage-point decline implies that users write substantially less toxic comments when pre-moderation is in place. That is, the reduction in toxicity among submitted comments accounts for 88% of the decline in initially published toxicity.

All the results presented in Figure 3 are robust to a wide range of robustness tests:

different cut-offs for the lag between publication and randomization (SI 6.1); alternative toxicity measures (SI 6.5); including controls (chosen by Lasso-based post-double selection, as pre-registered; see SI 6.2); not winsorizing outcomes (SI 6.3); excluding the most active commenters to make sure they don't drive results (SI 6.7); accounting for noncompliance with a LATE estimation (SI 6.8); and using seemingly unrelated regression to account for correlation among outcomes (SI 6.9).² These patterns indicate that the primary pathway through which pre-moderation improves discourse is by shifting what users write—rather than filtering content after submission.

2.2 Mechanisms

Our main results show that algorithmic pre-moderation reduces toxicity relative to post-moderation alone—and that its effects are largely driven by changes in user submission. This section shows that these effects are driven by behavioral change rather than selection, and arise from altered normative cues early in the discussion.

2.2.1 Selection or behavioral change?

The decline in toxic submissions to pre-moderated articles could arise through two alternative channels. First, pre-moderation may affect the composition of commenting users: less-toxic users may self-select into pre-moderated comment sections (or more-toxic users may self-select out). Second, the same users may submit different comments under alternative moderation regimes, adapting their behavior to match the prevailing tone.

No selection on user composition. SI Figure 25a plots group means for two user characteristics computed before the experiment: each user's pre-experiment average toxicity and their pre-experiment comment count. On both dimensions, the control and pre-moderation

²As pre-registered, we also estimate a specification in which we show treatment effects on all article comments, rather than just post-randomization comments (SI 6.4); as might be expected, including pre-treatment comments reduces the treatment effects.

means are essentially indistinguishable, indicating that users who later comment on treated versus control articles were similarly toxic (p-value of the difference in group means is $p = 0.349$) and similarly active before the experiment ($p = 0.230$). Covariate-adjusted models (SI Figure 25b) and non-winsorized specifications (SI Figures 26a-26b) reproduce the same null differences, reinforcing the absence of selective participation by less-toxic users.

Same users, different behavior. To isolate changes in behavior within user, we re-estimate our main regressions at the comment level, adding user fixed effects (SEs clustered and weighted by article, see SI 8.1.2; this analysis is exploratory). Adjusting for fixed effects, the within-user reduction is about 1.0 percentage point ($p = 0.004$), very similar to the article-level ITT. This shows that the same user writes less-toxic comments when articles are pre-moderated.

Taken together, these analyses strongly indicate that behavioral change by the same users, not in- or out-selection, is the primary mechanism driving pre-moderation’s reduction of toxicity: users do not sort into cleaner comment sections; they write more civilly when toxic cues are suppressed.

2.2.2 Why do users change their behavior?

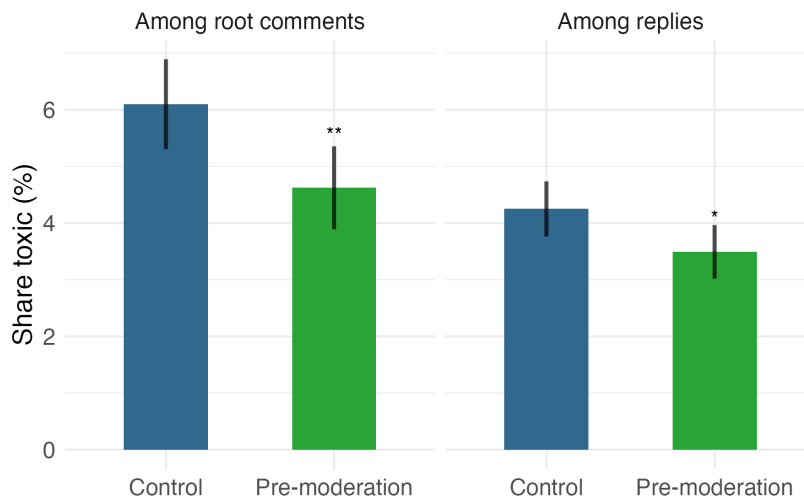
The evidence presented above raises the question why users write less toxic comments under pre-moderation. We consider three, non-exclusive channels:

Deterrence from perceived monitoring. Pre-moderated articles are not labeled or otherwise signaled to users, so they cannot directly observe whether a discussion is subject to pre-publication screening. Any adjustment in behavior based on expected monitoring would therefore have to be due to users inferring pre-moderation of an article based on indirect cues, such as the deletion or delay of their first post. If users infer that an article is pre-moderated and anticipate stricter enforcement, they may self-censor. We test this self-censoring mechanism in an exploratory analysis by restricting our sample to each user’s *first*

submitted comment on an article—made before the user can plausibly perceive delay due to pre-moderation. SI 8.4 shows that effects persist in this subsample and that pre-moderation decreases toxicity from 5.3% in control to 4.1% in treated group ($p = 0.004$) among users’ first comments. This suggests that deterrence alone cannot explain the observed change.

Reply-cascade. If toxicity is filtered among *root* comments, later replies are less likely to escalate [24]. We test this mechanism in an exploratory analysis by examining effects separately for root comments and replies; if the effect were driven primarily by the prevention of toxic reply cascades, we would expect stronger treatment effects among replies. However, Figure 4 shows that toxicity declines in both root comments (4.6% in treatment vs. 6.1% in control, $p = 0.008$) and replies (3.5% in treatment vs. 4.3% in control, $p = 0.029$), with the treatment effect among replies not always statistically significant across specifications (see SI 8.3). While the difference in treatment effects among roots and replies is not significant at conventional levels ($p=0.27$), the reduction is, if anything, larger and more robust among root comments, suggesting that halting toxic reply cascades is not the primary source of pre-moderation’s effectiveness.

Figure 4: Pre-moderation reduces toxicity most among both root comments and replies



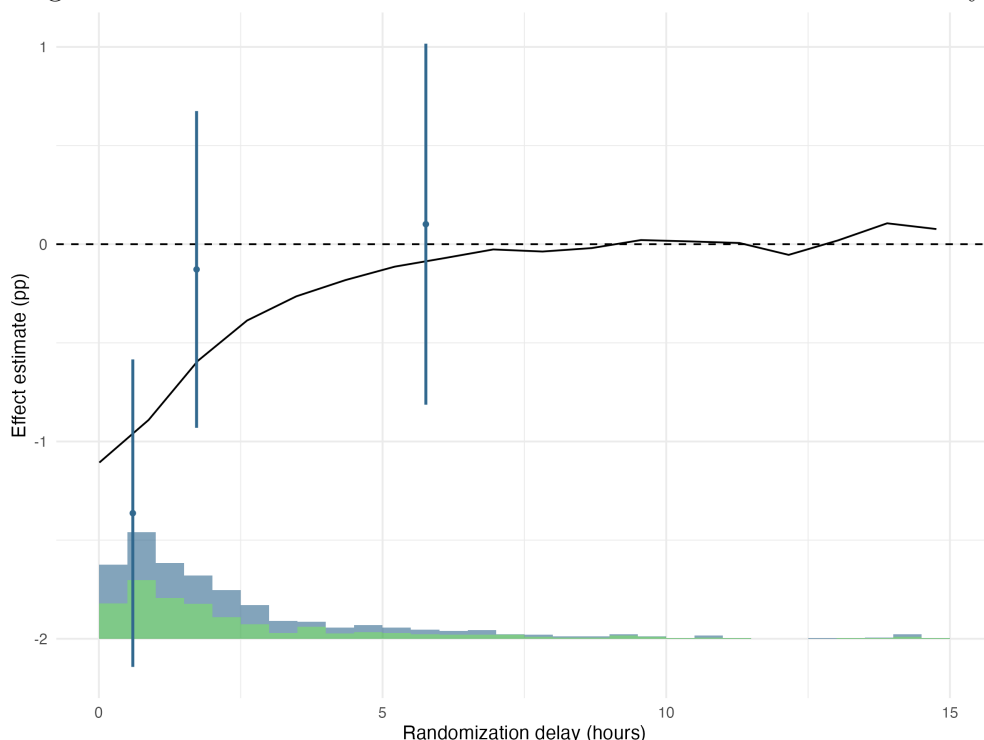
Note: +, *, **, *** denote significance of the differences in means at the 10%, 5%, 1% and 0.1% levels, respectively.

Tone-matching (normative cues). A third—and most strongly supported—mechanism is tone-matching in response to salient normative cues. Prior work shows that users adapt their language to the tone of visible content, using early contributions as signals of what behavior is appropriate in a given discussion space [10]. By filtering toxic comments before they become visible, pre-moderation alters the informational environment users encounter when deciding how to contribute. In doing so, it elevates the salience of civil norms and prevents toxic comments from establishing a hostile or confrontational tone [9].

Consistent with this mechanism, the previous two tests show that users write more civil first posts and root comments under pre-moderation—patterns that are difficult to reconcile with deterrence or reply-cascade explanations but follow directly from tone-matching driven by early normative cues. Further support for this mechanism comes from the timing of treatment exposure. Since the timing of randomization was not itself experimentally manipulated, we treat this analysis as exploratory despite its pre-registration. Figure 5 shows that earlier randomization of pre-moderation—shortly after article publication, before toxic comments have an opportunity to accumulate—is associated with substantially larger reductions in the toxicity of submitted comments. Splitting the sample into three equal-sized groups, we find that the benefits of pre-moderation are concentrated among articles randomized earliest, within 64 minutes of publication (-1.4pp , $p = 0.001$). In contrast, when randomization occurs later, after initial comments are already visible, effects are small and statistically indistinguishable from zero (-0.1pp , $p = 0.76$ for mid-delay articles; 0.1pp , $p = 0.83$ for high-delay articles; see SI Section 6.1.2 for details). Taken together, these results suggest that pre-moderation works primarily by shaping the normative environment at the outset of a discussion.

Building on the evidence for tone-matching, we study how toxicity evolves as the discussion unfolds. Using comments submitted before and after randomization, exploratory analysis in Figure 6 shows that group means are virtually identical pre-treatment and then diverge in the first 12 hours post-randomization: toxicity rises in control comment sections

Figure 5: Pre-moderation is most effective in articles randomized early



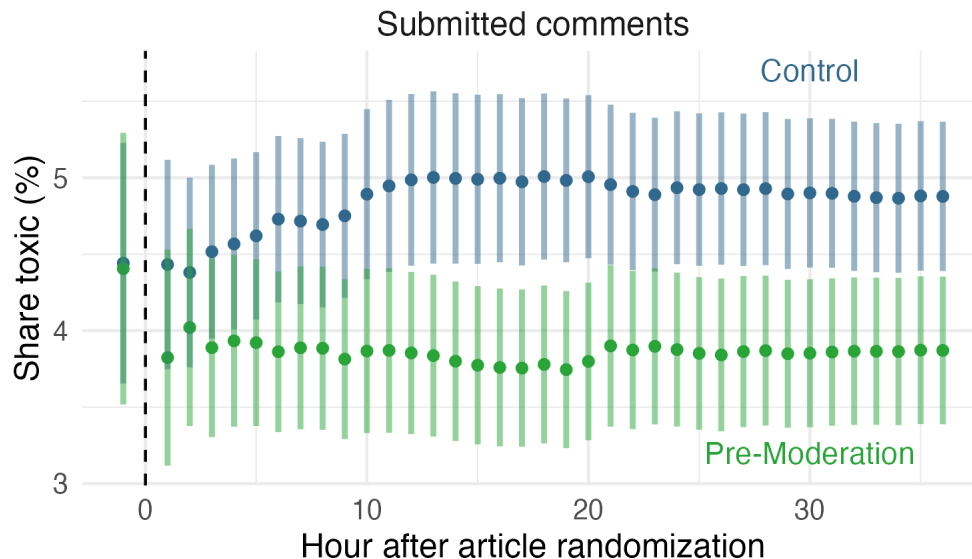
Note: The chart shows the estimated effect of pre-moderation on the share of toxic comments among submitted comments. The article sample is split into three equal-sized groups based on the delay between article publication and randomization. The first group encompasses delays below 63.79 minutes (median delay: 35.88), the second group encompasses delays between 63.79 and 161.72 minutes (median delay: 103.31), and the third group encompasses delays above 161.72 minutes (median delay: 345.78 minutes). The histogram shows the frequency distribution of pre-moderated articles (green) and control articles (blue), stacked within each delay bin.

as conversational norms erode, but remains flat in treated comment sections. Exploratory analysis presented in SI Figure 24 replicates this result for initially published and ultimately published comments. Together, these findings show that, absent pre-moderation, discussions tend to escalate in toxicity, whereas early removal of toxic content stabilizes discussions by preventing toxic cues from anchoring subsequent contributions and setting the tone.

2.2.3 Unbundling the treatment

In practice—and in our experiment—pre-moderation bundles two distinct components: (i) *pre-publication deletion* of comments flagged as harmful and (ii) *pre-publication review* of borderline comments, which introduces delay and downranking, as comments appear in the

Figure 6: Pre-moderation prevents discussions from escalating



feed in reverse chronological order by submission time rather than publication time. Consistent with this design feature, comments that undergo review but are eventually published are 4.5 percentage points less likely to receive any replies than comments published without review ($p < 0.001$), and take around 55 minutes longer before receiving their first reply ($p = 0.003$) (see SI Section 8.2).

Although pre-publication deletion and review were not randomized separately, two patterns suggest that deletion is the dominant channel driving the observed reductions in toxicity. First, comments that were initially deleted are substantially more toxic than those that were delayed but eventually published (21.4% versus 9.1%, $p < 0.001$; see SI Section 8.2)³, and a large share of reviewed comments are eventually published (77.7%). Second, SI Figure 27 shows that across outcomes—the toxic share among initially published, ultimately published, and submitted comments—articles in which pre-moderation operates primarily through deletion exhibit larger post-randomization reductions in toxicity than articles in which it operates mainly through review and downranking. Taken together, these exploratory patterns are consistent with our earlier mechanism tests and indicate that re-

³Of course, both are more toxic than those published without delay (4.2%).

moving toxic content before it can set the tone—rather than merely delaying or downranking it—is central to the effectiveness of pre-moderation.

2.3 Potential drawbacks of pre-moderation

A common concern is that pre-moderation depresses engagement. Prior work shows that users exposed to toxicity increase their social media activity [11, 25, 41]. Furthermore, in our setting, pre-moderation not only mechanically reduces the volume of published comments by filtering the most toxic, but also reduces the speed of ultimately published comments for the $\sim 5\%$ of comments that undergo manual moderation. Despite these pressures, we find no evidence of reduced engagement: treatment effects are positive and statistically insignificant.⁴ The average number of submitted comments per article is similar in treatment and control (278 vs. 262; $p = 0.64$, $p_{BH} = 0.64$), as is the number of unique commenters (131 vs. 125; $p = 0.70$, $p_{BH} = 0.70$; see SI Figure 29). Controlling for covariates, we can rule out that the intervention decreased engagement by more than 16.2 comments and 9.8 unique commenters, which correspond to small effects of just 0.07 and 0.05 standard deviations (one-sided tests with 95% confidence; see SI Section 9.1.2). Exploratory outcomes—time spent in the comments section and comments-page impressions—point in the same direction and do not suggest declines (see SI Section 9.1).

Another potential drawback concerns over-moderation, i.e. the deletion of content that should not have been removed [14, 34]. Our data and classifiers are limited in their ability to conclusively assess whether pre-moderation increases the risk of such deletions. As an exploratory check, we estimate the effect of pre-moderation on the share of non-toxic deleted comments among all submitted comments. We find a small increase in this share, from 5.4% under post-moderation to 6.0% with pre-moderation, which is not statistically significant ($p = 0.37$; exploratory analysis, see SI Section 9.2). Moreover, this measure likely overstates true over-moderation, as some deleted non-toxic comments violate other platform rules un-

⁴We did, however, find that—likely because some root comments were published with a delay—pre-moderation led to a shift from replies to root comments.

related to toxicity. Taken together, these exploratory results suggest that the reduction in toxicity achieved by pre-moderation comes, at most, at the cost of a modest increase in deletions unrelated to toxic content.

3 Discussion

Our research shows that algorithmic pre-moderation improves discourse through two reinforcing pathways. First, it prevents exposure to highly toxic comments at the outset, lowering the toxic share among both *initially* and *ultimately* published content. Second, and crucially, it shifts how people write: the same users submit less-toxic comments when pre-moderation is in place, with reductions for both root comments and replies. Indeed, 85% of the reduction in published toxicity is driven by changes in submissions.

In our setting, post-moderation removes little toxic content and does so only after exposure. The treatment–control gap in toxicity barely narrows from initial to ultimate publication, user reporting rates are low, and later deletions of published toxicity are rare. Because visibility precedes removal, toxic posts can still set the tone and invite mimicry before any enforcement occurs. Pre-moderation overcomes these limits by acting before publication and without relying on users to flag harmful content.

Public debates often frame moderation as a trade-off between free expression and the amount of content removed. Our results add an underappreciated dimension: removal policies also work *behaviorally*, through tone-matching. By curbing early toxic cues, pre-moderation alters local norms and induces more civil contributions from the same users. These multiplier effects help explain why discourse quality improves beyond the purely mechanical effect of filtering, and why this can be achieved without detectable losses in engagement.

Several features of our study constrain generalizability. First, the implementation of both pre-moderation and—especially—post-moderation is platform-specific, depending on

factors such as toxicity thresholds, moderator capacity, and how user reports are surfaced. *Der Standard* relies on a custom classifier for pre-moderation that is not publicly available; using an off-the-shelf tool may yield a different level of effectiveness. As larger platforms’ tools are likely more cutting-edge, they may be able to implement pre-moderation even more successfully. Post-moderation, which we find to be relatively ineffective at reducing toxicity in our setting, might perform better on platforms with more active moderators or users who report harmful content at higher rates [e.g. 42], though low reporting rates are a pervasive phenomenon [e.g. 43]. At the same time, our setting likely features comparatively strong community norms relative to many social media platforms and does not face the dynamics of virality, which can further impede post-moderation [18, 19]. Second, timing matters: some articles were randomized with delay, and we estimate larger effects when pre-moderation was activated earlier. This suggests that our estimates likely present a lower bound on the effects of pre-moderation implemented at the outset. Third, we evaluate an article-level intervention rather than platform-wide policy. As a result, most users were likely exposed to both moderation regimes during the experiment. To the extent that spillovers occur across articles, our estimates may understate the effects of a consistent platform-wide rollout—though such a rollout might also lead users to move to less moderated platforms [38]. Fourth, toxicity measurement relies on a validated but imperfect classifier trained on German- and French-language online comments from newspapers and social media. We therefore hope future work will replicate these tests on other platforms and in other countries, with different user populations, languages, classifiers, and post-moderation workflows. At the same time, there are reasons to expect broader applicability: the primary mechanism we document—tone-matching—is a well-established psychological behavior that should operate across contexts [9].

Our findings have implications for theory, platform design, and policymakers. For theory, the evidence that the same users become less toxic under pre-moderation supports accounts of *tone-matching* and the power of local descriptive norms in online settings. This suggests that

the design of digital platforms substantially affects users’ propensity of being toxic online, contradicting prior research that has treated online hostility as a stable individual character trait [e.g. 44]. When early toxic cues are suppressed, users align with more civil visible content, consistent with experimental evidence on normative influence in lab environments [9] and observational evidence from Reddit [10]. Together, these results emphasize that conversational context—what is visible when users decide what to write—is a key lever for steering norms, complementing prior work on anonymity and community standards [22, 23].

For platforms, two design lessons follow. First, our results on tone-matching and escalation dynamics suggest that platforms should act early and prioritize pre-publication screening for high-risk topics or time windows, and minimize delays between publication and moderation to prevent toxic tone-setting. Second, platforms should combine pre-moderation with complementary measures that reinforce civil norms (e.g., visible moderator presence) rather than relying solely on user flagging [45].

For policymakers, our results suggest that notice-and-action regimes centered on ex post removal are likely insufficient on their own. Pre-moderation combined with transparency about algorithmic screening is required. Lastly, this research also highlights how privacy-preserving data access for independent evaluation, as requested by the EU’s Digital Services Act, can provide much-needed scientific evidence for healthier digital public spheres [14, 46, 47].

Materials and Methods

This experiment was approved by the ETH Zurich Ethics Commission (IRB, project number EK-2024-N-53-A) and all our outcomes and statistical analyses, unless otherwise noted, were preregistered on OSF before we obtained the data. Additional preregistered and exploratory analysis is reported in the SI.

Data and sample

da

Platform and sample. We partner with the Austrian newspaper *Der Standard* to conduct a randomized field experiment to assess how pre-moderation affects commenting behavior. Based on unique users, *derStandard.at* was the third-largest online medium in Austria in 2025 [39]. Our partner provides us with access to all comments submitted on their site, including never-published and published-then-deleted comments. The field phase of the experiment ran from April 6, 2024 to March 26, 2025.

The newspaper’s moderation team decided which articles are eligible for the experiment; they excluded some articles for which they wanted to retain control over moderation settings (incl. liveblogs, obituaries, sudokus). While we pre-registered a sample of 600 randomized articles across the two groups, we became aware during the course of the experiment that some articles were being randomized many hours after publication. We thus increased the number of overall randomized articles and chose a time cut-off that maximizes our power: randomization within 95 minutes of publication, which gives us a sample of 410 articles: 194 in the treatment group (47.3%) and 216 in the control group (52.7%). See SI Section 6.1 for details and robustness to different cut-offs.

Outcomes. Our outcomes are measured at the comment level and aggregated to the article level, at which the intervention takes place.

- Toxicity (defined as threatening, insulting, discrediting, vulgar, humiliating, hateful, harmful, or exclusionary communication): (a) *share of **initially** (ever) published comments that are toxic*; (b) *share of **ultimately** published comments (after 5 days of post-moderation) that are toxic*; and (c) *share of **submitted** comments that are toxic*.

We measure comment toxicity using a custom BERT-based classifier, PubPol/Toxicity⁵.

⁵The model will be published on the Huggingface model repository alongside alongside the manuscript

The model was developed by fine-tuning a multilingual BERT architecture [48] on over 435,000 expert-annotated German- and French-language online comments in two steps, using first the Swiss Hate Speech Corpus [49] and then our own annotated dataset of over 14,000 German-language newspaper online comments, and conducted extensive hyperparameter searches (see SI Section 4.1.1). To validate this classifier for this study, research assistants annotated a test set of 1,600 comments from *Der Standard*, efficiently stratified following [50]. We use this test set to select the probability threshold that maximizes the F_1 score for our main analysis. At this 0.75 threshold, the classifier performed with an F_1 of 0.35 and a Bookmaker Informedness (BM) of 0.32, confirming its validity for this application. On our original data, the classifier outperforms Google’s Perspective API for toxicity (the longstanding industry standard), even when optimizing the threshold for the Perspective classifier ($F1_{Perspective} = 0.52$ versus $F1_{Finetuned} = 0.55$).

- Post-moderation (measured 5 days after publication): (a) *share of initially published comments that is reported for deletion by users*; (b) *share of initially published comments that are later deleted* (exploratory); (c) *share of submitted comments that are deleted but not classified as toxic* (exploratory).
- Engagement: (a) *number of comments submitted*; (b) *number of unique users* per article comments section.

Randomized treatment. The newspaper’s moderation team entered selected articles into a custom webapp which randomized articles with probability 0.5 to treatment and control condition, using simple randomization. SI Section 2.3 provides more detail on the webapp. The treatment consists of applying the newspaper’s default moderation classifier to screen comments before they are published. SI Section 2.4 provides more details on the classifier. In both treated and control articles, moderators review comments after publication, generally prioritizing comments that are reported by users for post-moderation; SI Section 2.2 describes

the newspapers’ moderation workflow.

SI Section 2.5 provides a manipulation check that confirms that the pre-moderation treatment indeed increases the likelihood of comments being deleted before publication (2.1% vs 1.4% in the control condition, $p = 0.006$) and manually checked before being posted (5.5% vs 2.5%, $p = 0.000$).

SI Table 3 in SI Section 3.3 reports that randomization successfully balanced covariates as expected. An omnibus balance test shows that overall, our 89 covariates are balanced across treatment and control (p-value of a χ^2 -test: 0.26). Moreover, our outcomes of interest are balanced in the period between article publication and randomization.

Statistical methods

Statistical analysis. All our statistical analysis is pre-registered. In our main analysis, we include only comments submitted *after randomization*, since comments submitted before randomization are not subject to the randomized moderation setting. Analysis using all comments (including pre-randomization) is reported in SI Section 6.4. We use (a) unadjusted OLS regressions and (b) a Lasso-based post-double selection model which automatically selects covariates based on their predictiveness for outcomes and treatment assignment [51]; see SI Section 3.2 for more details on the covariates. We report unadjusted estimates in the main text and the covariate-adjusted ones in SI Section 6.2; results are very similar across specifications. For all models, we use heteroscedasticity-robust standard errors. We adjust p-values using the Benjamini-Hochberg procedure [52]; we adjust p-values separately for primary, secondary (including primary), and exploratory (including primary and secondary) outcomes (see SI Section 1). We winsorize outcomes at the 2nd and 98th percentiles to deal with outliers; results are robust to not winsorizing, see SI 6.2-6.3.

Since we observe 38.4% non-compliance with the experimental protocol—articles in which moderators did not apply the setting randomized by the webapp—our OLS estimates measure an intention-to-treat effect. SI Section 6.8 presents the corresponding local average

treatment effects (LATE), which are slightly larger in magnitude.

Acknowledgments

We are grateful to the community team at *Der Standard* for working with us to design and implement this intervention, as well as for providing access and guidance to their comment data. We acknowledge financial support of the Swiss National Science Foundation (Grant 10001618).

Disclosure

Der Standard is owned by a company whose primary shareholders are L.B.’s father, brother, and mother.

References

- [1] Müller, K & Schwarz, C. (2018) Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association*.
- [2] Williams, M. L, Burnap, P, Javed, A, Liu, H, & Ozalp, S. (2020) Hate in the machine: Anti-black and anti-muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology* **60**, 93–117.
- [3] Bursztyn, L, Egorov, G, Enikolopov, R, & Petrova, M. (2019) Social media and xenophobia: evidence from russia, (National Bureau of Economic Research), Technical report.
- [4] Braghieri, L, Levy, R, & Makarin, A. (2022) Social media and mental health. *American Economic Review* **112**, 3660–3693.
- [5] Lorenz-Spreen, P, Oswald, L, Lewandowsky, S, & Hertwig, R. (2023) A systematic

- review of worldwide causal and correlational evidence on digital media and democracy. *Nature human behaviour* **7**, 74–101.
- [6] Mjelva, M. B, Schutte, S, Binningsbø, H. M, & Jensenius, F. R. (2025) Beyond hate speech: Online rumors and out-group resentment in divided societies. *Comparative Political Studies* p. 00104140251381753.
- [7] Kim, T. (2023) Violent political rhetoric on twitter. *Political science research and methods* **11**, 673–695.
- [8] Piccardi, T, Saveski, M, Jia, C, Hancock, J, Tsai, J. L, & Bernstein, M. S. (2025) Reranking partisan animosity in algorithmic social media feeds alters affective polarization. *Science* **390**, eadu5584.
- [9] Álvarez-Benjumea, A & Winter, F. (2018) Normative Change and Culture of Hate: An Experiment in Online Environments. *European Sociological Review* **34**, 223–237.
- [10] Rajadesingan, A, Resnick, P, & Budak, C. (2020) *Quick, community-specific learning: How distinctive toxicity norms are maintained in political subreddits*. Vol. 14, pp. 557–568.
- [11] Beknazar-Yuzbashev, G, Jiménez-Durán, R, McCrosky, J, & Stalinski, M. (2025) Toxic content and user engagement on social media: Evidence from a field experiment.
- [12] Dimmery, D, Bronner, L, Kotarcic, A, Gilardi, F, Donnay, K, Grech, P, & Hangartner, D. (2025) Quasi-experimental controlled direct effects: Measuring the effects of a moderation system. Working Paper.
- [13] Einwiller, S. A & Kim, S. (2020) How online content providers moderate user-generated content to prevent harmful online communication: An analysis of policies and their implementation. *Policy & Internet* **12**, 184–206.

- [14] Kozyreva, A, Herzog, S. M, Lewandowsky, S, Hertwig, R, Lorenz-Spreen, P, Leiser, M, & Reifler, J. (2023) Resolving content moderation dilemmas between free speech and harmful misinformation. *Proceedings of the National Academy of Sciences* **120**, e2210666120.
- [15] Hartmann, D, Oueslati, A, Staufer, D, Pohlmann, L, Munzert, S, & Heuer, H. (2025) *Lost in moderation: How commercial content moderation APIs over-and under-moderate group-targeted hate speech and linguistic variations*. pp. 1–26.
- [16] Srinivasan, K. B, Danescu-Niculescu-Mizil, C, Lee, L, & Tan, C. (2019) Content removal as a moderation strategy: Compliance and other outcomes in the changemyview community. *Proceedings of the ACM on Human-Computer Interaction* **3**, 1–21.
- [17] Wirtschafter, V & Majumder, S. (2023) Future challenges for online, crowdsourced content moderation: evidence from twitter’s community notes. *Journal of Online Trust and Safety* **2**.
- [18] Chuai, Y, Tian, H, Pröllochs, N, & Lenzini, G. (2024) Did the roll-out of community notes reduce engagement with misinformation on x/twitter? *Proceedings of the ACM on Human-Computer Interaction* **8**, 1–52.
- [19] Goldstein, I, Edelson, L, Nguyen, M.-K, Goga, O, McCoy, D, & Lauinger, T. (2023) Understanding the (in) effectiveness of content moderation: A case study of facebook in the context of the us capitol riot. *arXiv preprint arXiv:2301.02737*.
- [20] Meta Platforms, Inc. (2026) Fourth quarter and full year 2025 earnings conference call (Earnings conference call and press release). Press release available at <https://investor.atmeta.com/investor-news/press-release-details/2026/Meta-Reports-Fourth-Quarter-and-Full-Year-2025-Results/default.aspx>. Transcript PDF available at <https://investor.atmeta.com>.
- [21] Kaplan, J. (2025) More speech and fewer mistakes (Meta Newsroom).

- [22] Santana, A. D. (2014) Virtuous or vitriolic: The effect of anonymity on civility in online newspaper reader comment boards. *Journalism practice* **8**, 18–33.
- [23] Hille, S & Bakker, P. (2014) Engaging the social news user: Comments on news sites and facebook. *Journalism Practice* **8**, 563–572.
- [24] Ribeiro, M. H, Cheng, J, & West, R. (2023) Automated content moderation increases adherence to community guidelines.
- [25] Jiménez Durán, R. (2021) The economics of content moderation: Evidence from hate speech on twitter. *Available at SSRN 4044098*.
- [26] Habibi, M, Hovy, D, & Schwarz, C. (2024) The content moderator’s dilemma: Removal of toxic content and distortions to online discourse. *arXiv preprint arXiv:2412.16114*.
- [27] Jhaver, S, Bruckman, A, & Gilbert, E. (2019) Does transparency in moderation really matter? user behavior after content removal explanations on reddit. *Proc. ACM Hum.-Comput. Interact.* **3**.
- [28] Myers West, S. (2018) Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society* **20**, 4366–4383.
- [29] Ma, R, You, Y, Gui, X, & Kou, Y. (2023) How do users experience moderation?: A systematic literature review. *Proceedings of the ACM on Human-Computer Interaction* **7**, 1–30.
- [30] Cosley, D, Frankowski, D, Kiesler, S, Terveen, L, & Riedl, J. (2005) *How oversight improves member-maintained communities*. pp. 11–20.
- [31] Wise, K, Hamman, B, & Thorson, K. (2006) Moderation, response rate, and message interactivity: Features of online communities and their effects on intent to participate. *Journal of Computer-Mediated Communication* **12**, 24–41.

- [32] Pan, C. A, Yakhmi, S, Iyer, T. P, Strasnick, E, Zhang, A. X, & Bernstein, M. S. (2022) Comparing the perceived legitimacy of content moderation processes: Contractors, algorithms, expert panels, and digital juries. *Proceedings of the ACM on Human-Computer Interaction* **6**, 1–31.
- [33] Pradel, F, Zilinsky, J, Kosmidis, S, & Theocharis, Y. (2024) Toxic speech and limited demand for content moderation on social media. *American Political Science Review* **118**, 1895–1912.
- [34] Munzert, S, Traunmüller, R, Barberá, P, Guess, A, & Yang, J. (2025) Citizen preferences for online hate speech regulation. *PNAS Nexus* p. pgaf032.
- [35] Katsaros, M, Yang, K, & Fratamico, L. (2021) Reconsidering tweets: Intervening during tweet creation decreases offensive content.
- [36] Allcott, H, Gentzkow, M, Mason, W, Wilkins, A, Barberá, P, Brown, T, Cisneros, J. C, Crespo-Tenorio, A, Dimmery, D, Freelon, D, et al. (2024) The effects of facebook and instagram on the 2020 election: A deactivation experiment. *Proceedings of the National Academy of Sciences* **121**, e2321584121.
- [37] Guess, A. M, Malhotra, N, Pan, J, Barberá, P, Allcott, H, Brown, T, Crespo-Tenorio, A, Dimmery, D, Freelon, D, Gentzkow, M, et al. (2023) How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science* **381**, 398–404.
- [38] Kalra, A. (2025) Hate in the time of algorithms: Evidence on online behavior from a large-scale experiment. *arXiv preprint arXiv:2503.06244*.
- [39] Österreichische Web-Analyse. (2025) Reichweitenstudie 2025.
- [40] Public Discourse Foundation. (2026) Public Discourse Indicator (<https://www.public-discourse.org/en/public-discourse-indicator>). Accessed: 2026-02-01.

- [41] Oswald, L, Schulz, W. S, & Lorenz-Spreen, P. (2025) Disentangling participation in online political discussions with a collective field experiment. *Science Advances* **11**, eady8022.
- [42] Schneider, P. J & Rizoiu, M.-A. (2023) The effectiveness of moderating harmful online content. *Proceedings of the National Academy of Sciences* **120**, e2307360120.
- [43] Anti-Defamation League. (2024) Online hate and harassment: The american experience 2024.
- [44] Bor, A & Petersen, M. B. (2022) The psychology of online political hostility: A comprehensive, cross-national test of the mismatch hypothesis. *American political science review* **116**, 1–18.
- [45] Bronner, L, Tomas-Valiente, F, Berk, N, & Hangartner, D. (2025) Tackling harmful online comments with visible moderator presence. Working paper.
- [46] European Parliament and Council of the European Union. (2022) Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act) (Official Journal of the European Union L 277, 27 October 2022, pp. 1–102). OJ L 277, 27.10.2022.
- [47] Parliament of the United Kingdom. (2023) Online safety act 2023 (Public General Acts). Royal Assent 26 October 2023.
- [48] Devlin, J, Chang, M, Lee, K, & Toutanova, K. (2018) BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR* **abs/1810.04805**.
- [49] Kotarcic, A, Hangartner, D, Gilardi, F, Kurer, S, & Donnay, K. (2022) Human-in-the-loop hate speech classification in a multilingual context. *arXiv preprint arXiv:2212.02108*.

- [50] Tomas-Valiente, F. (2025) Uncertain performance: How to quantify uncertainty and draw test sets when evaluating classifiers. *Working paper*.
- [51] Belloni, A, Chernozhukov, V, & Hansen, C. (2014) Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies* **81**, 608–650.
- [52] Benjamini, Y & Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **57**, 289–300.