Proceedings of the Twentieth International AAAI Conference on Web and Social Media (ICWSM 2026)

Shifting Gatekeeping Practices and Audience Engagement Under AI-Assisted Comment Moderation

Ashwin Rajadesingan
Department of Communication Studies
University of Texas at Austin
arajades@austin.utexas.edu

Abstract

Traditionally, The New York Times has maintained substantial gatekeeping control over audience contributions on its website by selectively enabling comment sections on certain articles and pre-moderating comments. This study explores changes to these gatekeeping practices and subsequent audience engagement following the transition to AI-assisted comment moderation. The introduction of AI-assisted comment moderation coincided with an increase in articles with comment sections. However, we found that this increase was less pronounced for contentious, social issue topics than for soft news sections. Comparing comment quality before and after AI assistance across eight dimensions such as coherence and substantiveness, we observed some section-wise differences, although overall, the differences were minimal. We also found that AI-assisted moderation was associated with an increase in engagement among prolific commenters and an increase in the proportion of moderately active commenters. Overall, these findings indicate that AI-assisted moderation is associated with changes in gatekeeping and increased audience participation without compromising on the aggregate quality of comments published.

1 Introduction

In online news sites, a key application of AI is in moderating comments on news articles. With readers contributing thousands of comments daily, many online news sites have adopted a hybrid approach combining human moderators and AI to scale up comment moderation (Weisner, Schäfer, and Lecheler 2025). In this study, we examine how the adoption of AI-assisted comment moderation by news organizations is associated with changes in their traditional gatekeeping practices and subsequent audience engagement.

For online news sites, comment sections present a curious dilemma. In theory, comment sections could serve as productive discussion spaces furthering the ideals of deliberative democracy (Bächtiger et al. 2018). These spaces may also increase engagement with news content helping drive readership (Engelke 2020). However, in practice, comment sections have often devolved into toxic spaces hosting bitter, hostile exchanges with limited constructive dialogue (Watson, Peng, and Lewis 2019). Further, research suggests that the incivility in news comments sections may negatively skew the perception of the journalistic content they read on the sites (Masullo, Tenenboim, and Lu 2023).

Some news sites including The New York Times (NYT) and The Guardian, have responded to this predicament by (i) restricting commenting to select articles and (ii) employing professional moderators, enabling them to focus their limited moderation resources where they are needed the most. The practice of selectively opening up comment sections for certain news articles and moderating them may be viewed as a Web 2.0 extension of the traditional gatekeeping functions that news outlets perform (Santana 2016; Juarez Miro 2022). News organizations function as gatekeepers of information. They exercise significant editorial judgment in selecting which events to report on given their limited resources and unlimited number of potential stories (Shoemaker and Vos 2009). Through gatekeeping, news media outlets steer attention towards and away from certain events, influencing the social reality perceived by their audiences (Soroka 2012). Similar to gatekeeping by news selection, online news sites may channel discussion towards and away from certain topics by selectively restricting comment sections to certain articles. Likewise, news sites also practice gatekeeping through comment moderation. News sites curate their comments section by deciding on which comments to retain or promote and which comments to remove, again directing attention towards and away from certain views (Juarez Miro 2022).

More recently, with advancements in AI, many news outlets have enabled AI-assisted comment moderation on their sites. For example, in 2017, the NYT in collaboration with Jigsaw, a Google-backed incubator, launched AI-assisted comment moderation which allowed them to transition from a fully human moderated process to a more hybrid approach (Etim 2017). This shift in approach may be associated with changes in media gatekeeping logic and associated patterns of audience engagement. Therefore, we conducted a large-scale analysis of comment sections on NYT before and after their adoption of AI-assisted content moderation. We summarize the results here:

1. The adoption of AI-assisted comment moderation was associated with a change in gatekeeping logic but not in the manner we expected. AI adoption significantly increased the frequency of articles with comment sections.

However, the largest increases occurred, not on articles about social issues but in largely soft news sections.

2. Overall, we found minimal differences in terms of quality measures in comments published before and after the AI-assisted comment moderation. However, there were some news-section specific differences especially in measures such as inflammatory content suggesting efforts to reduce potentially heated discussions in certain news sections. Thus, we found some evidence of changes in gatekeeping through comment moderation.
3. Regarding audience engagement, AI adoption corresponded to (i) an increase in participation among already prolific commenters (though there was little increase in participation among less prolific commenters) and (ii) a significant increase in the proportion of moderately active commenters. These results suggest that AI adoption was associated with changes not only in overall participation but also in how engagement was distributed among commenters.

2 Background

2.1 Gatekeeping audience generated content

One fundamental function of news media organizations is to gatekeep information. Shoemaker et al. (2001) explain gatekeeping as "the process by which the vast array of potential news messages are winnowed, shaped, and prodded into those few that are actually transmitted by the news media." According to the hierarchy of influences model (Shoemaker and Reese 2013), the gatekeeping process is shaped by factors operating at five levels: individual, communication routines, organizational, social institutional, and broader social system levels. Thus, the content published by news organizations is a result of this extensive, carefully refined gatekeeping process. Unlike news content, comment sections, which are typically placed right below the article text are almost entirely audience produced. In the comment sections, apart from providing their own commentary, users take on the role of gate-watchers with the ability to observe and rate comments controlling their visibility (Bruns 2021). Thus, by hosting audience generated content in the form of news comments, news organizations shift some of their gatekeeping control of information that is published on their platform to their audiences.

Nevertheless, news organizations have implemented strategies, akin to gatekeeping processes for news selection, to uphold journalistic norms such as civility and accuracy in the comments section (Hermida and Thurman 2008). One strategy to manage audience generated content is to open up only certain articles for comments. For example, The Guardian does not typically open comments for articles about "particularly divisive or emotional issues" or if it is "likely to create a hostile or negative discussion."[1] By allowing comments on select articles, news organizations are able to gatekeep certain information away from audience discourse. For example, research suggests that news sites are less likely to open articles on controversial topics for commenting due to concerns about managing potential incivility stemming from discussions on those topics (Santana 2016).

News organizations also shape audience contributions in comment sections through comment moderation (Boberg et al. 2018). Similar to the five-level gatekeeping process for news selection, researchers have identified a range of factors at different levels influencing comment moderation (Paasch-Colberg and Strippel 2022). Some organizations pre-moderate audience comments, that is, in-house moderators review and approve each submission before it is published. NYT applies pre-moderation for all its comments sections while others such as the Guardian pre-moderate certain sensitive articles. Pre-moderation allows significant control over audience discourse and most aligns with the traditional gatekeeping role that news organizations perform to select news articles (Hermida and Thurman 2008). Most organizations typically post-moderate most of their audience comments, that is, moderators review only comments flagged by other users. While post moderation requires fewer resources, it also limits the extent of gatekeeping by relying on users to flag perceived problematic content. The divergence in values that audiences prioritize in comment sections and the standards upheld by journalists can result in under-moderation of content and associated challenges (Juarez Miro 2022).

In this study, we examine how the introduction of AI-assisted comment moderation is associated with two key forms of gatekeeping practices exercised by news organizations: the selective opening of comment sections and the moderation of the comments themselves.

2.2 AI-assisted content moderation

A growing body of research on AI-assisted content moderation has studied how professional moderators, such as those working at NYT, interact with AI-based moderation systems. Qualitative studies based on interviews with moderators indicate a strong preference for a human-in-the-loop approach, in which AI systems provide decision support rather than autonomously determining moderation outcomes (Koelmann et al. 2022). Reflecting the inherent complexity of moderation work, moderators further emphasize the importance of AI systems with features that support collaborative decision-making with peers, offer transparent and explainable AI judgments, and allow moderators to override or correct AI-generated decisions (Niemann 2021). AI adoption for content moderation is not without moderator skepticism. Weisner et al. (2025) found that although moderators desired advanced tools to identify and remove problematic content, they were skeptical of its use for "gray" area comments that required nuanced understanding or specific contextual knowledge. Similarly, Ruckenstein et al. (2020) found that moderators were concerned about how the AI cannot keep up with newer forms of removable content that it has not been trained on previously, limiting AI use. McInnis et al. (2021) observe that AI tools enable moderators to scale their efforts by effectively detecting clearly removable content, while still requiring manual review for false negatives that slip through. These studies also highlight moder-

[1] https://www.theguardian.com/community-faqs

ators' desire for the AI to take over routine tasks of identifying problematic content to free them to focus on more interactive moderation approaches such as example setting that can shape and nurture the platform's discussion culture (Ruckenstein and Turunen 2020; Weisner, Schäfer, and Lecheler 2025; McInnis et al. 2021; Waterschoot 2024).

At the same time, moderators are acutely aware of the potential for bias in AI-assisted moderation systems (Salganik and Lee 2020) and seek greater transparency in automated decision-making (McInnis et al. 2021). Udupa et al. (2023) critique such systems for their limited ability to detect culturally coded forms of vitriol and extreme speech, particularly in contexts in the Global South. As an alternative to corporate-driven scaling, they propose "ethical scaling", which emphasizes community participation, transparent iteration, and a commitment to justice over profit-making. Similarly, in their analysis of the Perspective API, Rieder and Skop (2021) argue that AI moderation systems encode moral and cultural assumptions about what constitutes a "good" contribution into technical classifications, raising concerns about normative homogenization as these tools are adopted at scale. They further argue that accountable content moderation requires not only auditing algorithms, but also scrutinizing the institutional collaborations—(such as those between NYT and Jigsaw) that shape how human judgment, organizational values, and AI systems are jointly produced and deployed.

Prior research on the impact of AI-assisted comment moderation has largely focused on users' perceptions of credibility (Molina and Sundar 2022; Wang 2023) and quality of content produced (Horta Ribeiro, Cheng, and West 2023; Mohammadi and Yasseri 2025; Yu et al. 2024). By contrast, little attention has been paid to how AI-assisted moderation is associated with changes in participation frequency and the distribution of participation across user subgroups, which is a core focus of this study.

2.3 Comment moderation at the New York Times

Perhaps reflecting its prominence in the field, the NYT comments section has been extensively studied by researchers to better understand news deliberation (Muddiman and Stroud 2017), audience participation (Pierson 2015) and moderation practices (Diakopoulos 2015; Juarez Miro 2022; He et al. 2024; Wang and Diakopoulos 2022; McInnis et al. 2021) in online news discourse. Relevant to this study, we discuss research that on comment moderation practices.

NYT has historically used pre-moderation by professional community managers, making moderation an extension of newsroom labor rather than simple content removal (Diakopoulos 2015). Moderator interviews show that the most consequential work occurs within the first 24 hours after comments open, when moderators prepare by closely reviewing the article, elevate exemplar comments to model norms, reject off-topic or toxic submissions, fact-check reader claims, and synthesize emerging discussion themes (McInnis et al. 2021). They also route reader critiques to appropriate newsroom stakeholders, acting as "secondary gatekeepers" who both regulate discourse and facilitate institutional accountability (McInnis et al. 2021).

This model of tight editorial control over comments section often highlights how NYT and its readers diverge in assigning value to public discourse, revealing competing logics. This tension is theorized as a "comment gap", where journalists elevate conciliatory, articulate, and diverse comments, while readers reward direct, confrontational, and ideologically aligned ones (Juarez Miro 2022). Large scale analyses confirm this observation, finding that comments containing profanity and partisan language receiving more user "recommendations" but are less likely to be highlighted as exemplars ("NYT Picks") (Muddiman and Stroud 2017). However, researchers argue that this apparent misalignment is a deliberate norm-setting moderation strategy (Wang and Diakopoulos 2022). Observational evidence indicates that receiving a Pick is associated with improved subsequent comment quality and increased commenting frequency (Wang and Diakopoulos 2022). Complementing this, work on the selection criteria for NYT Picks consistently finds that these featured comments are longer, more positive, more topically relevant and less toxic than non-featured comments, with some variation across newsroom sections (He et al. 2024; Diakopoulos 2015). We add to research on the NYT comments section by evaluating changes to comment quality after the introduction of AI-assisted comment moderation.

3 Hypothesis and Research Questions

The hypothesis and research questions can be categorized into two broad themes: how AI-assisted comment moderation is associated with (i) the decisions that media organizations make (gatekeeping) and (ii) how readers respond to those changes (audience engagement).

First, we evaluate gatekeeping when media outlets selectively open up news articles for commenting. Research suggests that a central concern for news organizations is ensuring that the comment sections are civil as the toxicity in comments adversely affects their brand identity and reduces audience participation (Meltzer 2015; Masullo, Tenenboim, and Lu 2023). Indeed, numerous major news sites have disbanded their comments section entirely citing the significant moderation resources needed to maintain decorum (Jensen 2016). Therefore, we expect fewer articles on social issues to have comment sections, because of their propensity to generate uncivil discussion (Santana 2016). We test the following:
H1: Articles about social issues are less likely to have comment sections compared to other articles.

The introduction of AI-assisted comment moderation may alter the aforementioned logic of gatekeeping to preserve civility in audience discourse. With its ability to scale up moderation efficiently, incivility may be less of a factor in opening up comment sections under articles about social issues. However, there is significant distrust among professional content moderators on the use of AI for moderation which may limit its impact (Weisner, Schäfer, and Lecheler 2025). Further, there may be other reasons to not scale up discussions on certain sensitive matters such as terrorist attacks as news organizations may want to retain control over framing and interpretation by limiting audience input (Ihlebæk and Krumsvik 2015). Therefore, we evaluate:

RQ1a: How does the availability of comment sections for social and non-social issue articles differ before and after the adoption of AI-assisted moderation?

The NYT organizes their content by news sections. We expect there to be some variability in the ability to comment on articles by news section. Certain news sections such as *Fashion* and *Theater* may be considered soft news while others such as *U.S.* and *World* may be considered hard news. While conceptual ambiguities persist, Reinemann et al. (2012) distinguish between soft and hard news simply as follows: "The more a news item is politically relevant, the more it reports in a thematic way, focuses on the societal consequences of events, is impersonal and unemotional in its style, the more it can be regarded as hard news. The more a news item is not politically relevant, the more it reports in an episodic way, focuses on individual consequences of events, is personal and emotional in style, the more it can be regarded as soft news." As these vastly different kinds of news sections may be associated with potentially different moderation standards (McInnis et al. 2021), we ask:

RQ1b: How does the availability of comment sections vary by news section before and after the adoption of AI-assisted moderation?

Second, we evaluate gatekeeping through comment moderation. The transition to AI-assisted comment moderation changed how reader comments were moderated. While previously, all submissions were manually reviewed by in-house community moderators, after the AI adoption, the comments were first run through an AI trained on previous moderation decisions on NYT comments. These changes may inadvertently affect the quality of comments published, potentially reconfiguring the gatekeeping process. Therefore, we evaluate:

RQ2a: How does the quality of comments published vary before and after the adoption of AI-assisted moderation?

Unlike RQ1a and RQ1b, where we have data on all published articles as well as the subset selected for commenting, our gatekeeping analysis for comment moderation is constrained by limited data availability. We do not have access to all comments submitted to the NYT, only those that were ultimately published. As a result, our analysis is restricted to published comments only. We further discuss the implications of this limitation in the limitations section. Given potential news section-wise variations in how the comments are moderated (McInnis et al. 2021), we also ask:

RQ2b: How does the quality of comments published vary by news section before and after the adoption of AI-assisted moderation?

The final set of RQs concern audience engagement. In theory, opening up more articles for commenting ought to increase reader engagement as this opens up more opportunities for commenting. However, in practice, the vast majority of users in social sites seldom have limited attention or interest to take even relatively less costly actions like up/-down voting (Gilbert 2013). Therefore, we evaluate:

RQ3a: How does the frequency of user commenting vary before and after the adoption of AI-assisted moderation?

Most commenting systems have skewed participation curves as most comments are made by a small set of prolific commenters (He et al. 2020), thus the effect of opening up more articles for commenting may have differential effects on prolific versus less active commenters. Therefore, we evaluate:

RQ3b: Are there differences in the frequency of user commenting between different commenter subgroups before and after the adoption of AI-assisted moderation?

Beyond differences in commenting frequency across subgroups, the relative size of these subgroups may also differ before and after the introduction of AI-assisted moderation. Considering both relative subgroup size and participation frequency offers a more complete picture of engagement changes. Note that for RQ3c, we examine the proportion of different commenter groups rather than the raw subgroup size to account for increases in overall NYT subscribership across the analysis period. Therefore, we assess:

RQ3c: Are there differences in the proportion of different commenter subgroups (based on the their frequency of commenting) before and after the adoption of AI-assisted moderation?

Note that for RQ3c, we examine the proportion of different commenter groups rather than the raw subgroup size to account for increases in overall NYT subscribership across the analysis period.

4 Why Study The NYT Comment Section?

We chose to study The New York Times because it is widely regarded as the US "newspaper of record". As a hugely influential news organization, NYT's decisions on digital strategies are reviewed by other outlets. Specific to moderation, NYT has a dedicated team of human moderators who moderate comments before they are published (pre-moderation). As a result, its comments section are widely considered to be of higher quality (O'Brien 2017).

The introduction of AI changed the in-house moderation process (Etim 2017). Previously, all submissions were reviewed by in-house moderators, but after AI was introduced, comments were first screened by a model trained on past NYT moderation decisions. The new moderation process was as follows[2]: for each comment submitted by a reader, the AI returned a probability indicating its likelihood of rejection by the NYT moderation team. Then, the comments were visualized as dots in a histogram based on this probability. The model provides information on why a comment was likely to be rejected (e.g. inflammatory or insubstantial). The moderators manually review a sample of comments from different sections of the histogram to determine if the comments ought to be published on or not.

5 Analysis Plan

On September 20, 2016, NYT announced a partnership with Jigsaw to build automated models to aid in comment moderation using their moderation decisions on past comments.[3] In the press release, they indicated that moving forward they

[2]https://www.nytimes.com/2017/06/13/insider/have-a-comment-leave-a-comment.html

[3]https://www.nytimes.com/interactive/2016/09/20/insider/approve-or-reject-moderation-quiz.html

would integrate AI-assisted comment moderation for some articles before a larger site wide launch. On June 13, 2017, NYT announced the public launch of their AI-assisted moderation process.[4] We consider the period between September 2016 and July 2017 to be a test and transition period for AI-assisted moderation and exclude those months from our analysis. Therefore, to conduct our analyses, we considered articles published during the 3 years prior to September, 2016 (pre-AI) and 3 years after July, 2017 (post-AI). The advantage of this set up is that we are able to align the pre-AI and post-AI time periods with their corresponding election cycles, that is both the pre-AI time period (August 2013 - August 2016) and post-AI time period (August 2017 - August 2019) start in the second year of the corresponding presidencies and end right before the elections. Aligning these time periods is crucial as research suggests that engagement with news and politics increase significantly in the run up to the elections when candidates campaign aggressively (León, Vermeer, and Trilling 2023). For our analyses, we first randomly sampled 12 months from the pre-AI time period. Then, for each randomly selected month in the pre-AI period, we selected its corresponding month four years later in the post-AI period. Table 5 in the Appendix lists all the pre-AI and post-AI months selected.

6 Data Collection

We scraped comments from articles published on the NYT website during months selected for analyses. For each selected month, we also obtained comments posted on articles published in the 6 months prior to that month. We used the NYT Archive API to obtain metadata on the articles published during those selected months. We only included articles that the NYT metadata identified as news and excluded blogs from the analyses. We also excluded articles in sections such as *Briefings* that do not have comment sections. We also excluded sections such as *Obituaries* that have fewer than 100 articles published during the 24 months that we analyzed. In total, we analyzed 79,824 articles of which 19,090 (23.92%) had an open comments section with a total of 3,603,990 comments. The Institutional Review Board (IRB) at the University of Texas at Austin determined that this study did not constitute human subjects research. However, given users' potential contextual expectations of privacy, we publicly release (i) complete article metadata (containing no user-generated content), and (ii) only the anonymized comment data necessary for replication.[5]

7 Measures

7.1 Identifying social issues articles

Since H1 and RQ1 focus on articles about social issues, we devised a reliable method to determine whether an article addresses a social issue or not. To do this, we relied on the keywords assigned to each article by NYT and included in the article meta data information provided by their NYT Archive API. These keywords, drawn from the NYT's controlled vocabulary, are used to classify articles across different products, including the Times Topics pages.

First, we identified 11 major social issues based on Pew Research Center surveys (Center 2020, 2016). While not exhaustive, these issues—such as abortion and gun rights—are known to frequently spark contentious online discourse (see Table 6). Following an approach similar to (Muddiman and Stroud 2017), we manually selected all potential keywords that aligned to each issue and appeared in at least 10 articles in our dataset. Our strategy was intentionally inclusive at this stage, prioritizing breadth over precision to avoid prematurely excluding relevant articles.

Then, we evaluated whether the articles associated with these keywords were truly relevant to the intended social issue. For each article, we extracted the title, lead paragraph, and abstract from the NYT Archive API and paired this information with the candidate topic. We then used OpenAI's `gpt-4.1-mini` model to classify each article as relevant or not to the social issue, also collecting the model's reasoning. The exact prompt used for this task is provided in the Appendix. All articles relevant to a social issue were classified as a social issue article.

We validated the model's performance through human annotation. Two annotators reviewed a random sample of 100 keyword-matched articles, determining whether each article was relevant to the model-assigned social issue (Krippendorff's $\alpha = 0.801$). After resolving disagreements, we compared the human labels to the model's outputs. The model achieved an accuracy of 90%, with a precision of 91.66% and a recall of 88%, performance metrics that apply specifically to articles pre-filtered by the keyword mapping step.

To further test the robustness of our keyword selection, we drew a separate random sample of 100 articles that *did not match* any of the selected keywords. The same annotators assessed whether any of these articles addressed one of the 11 social issues. Only 1 out of 100 articles in this sample was deemed to be about a social issue, suggesting that our keyword set had high coverage and few false negatives. In total, through this approach, we labeled 8,850 (11.09%) out of the 79,824 articles as being about a social issue.

To get a better understanding of the kind of articles selected under each topic, we include all keywords for which the model identified at least 60% of the articles mapped to them as relevant in Table 6 in the Appendix.

7.2 Measuring the quality of comments

Jigsaw made public the text classifiers trained solely on NYT comments using labels provided by the NYT moderation team. While it is unclear if these public models are the ones being used by NYT for moderation, they provide a viable public alternative to evaluate the quality of NYT comments "relying on patterns from more than 16 million acceptances and rejections of comments by Times moderators."[6] The classifiers provide the following measures of

[4] https://www.nytimes.com/2017/06/13/insider/have-a-comment-leave-a-comment.html

[5] https://github.com/Ashwin-R/icwsm2026-nyt-ai

[6] https://www.nytimes.com/2021/10/26/insider/why-humans-not-machines-make-the-tough-calls-on-comments.html

quality [7]: (i) likely to reject: likelihood of the comment being rejected by NYT's moderation (ii) attack on author: attack on the author of the article or post, (iii) attack on commenter: attack on fellow commenter, (iv) incoherent: difficult to understand, (v) inflammatory: intending to provoke or inflame, (vi) obscene: using obscene or vulgar language, (vii) spam: irrelevant and unsolicited commercial content, and (viii) unsubstantial: trivial comments and (ix) toxicity: a rude or unreasonable comment that is likely to make people leave a discussion. Before using these classifiers for analysis, we validated them on our dataset. We randomly sampled 100 comments each in the time period before and after AI-assisted comment moderation. To make precise comparisons of quality, it is important that (i) the classifiers perform adequately on our dataset and (ii) the accuracy of the classifiers remains consistent before and after introduction. To perform this validation, two annotators independently annotated the 200 sampled comments on eight dimensions [8], before discussing to resolve differences in labeling. The overall inter-rater agreement score (Krippendorff's α) was 0.56 which is inline with agreement levels on coding subjective concepts with high class imbalances (Wong, Paritosh, and Aroyo 2021) as most comments published are of high quality. We include Table 2 that shows the accuracy of the classifiers across dimensions on comments published before and after AI-assistance. The accuracy was calculated comparing classifier output to the consensus human labels. Note that while the accuracy is not uniform across dimensions, crucially, the accuracy for each dimension remains fairly consistent before and after AI-assisted moderation. Therefore, using these measures, we evaluated the quality of comments published before and after the adoption of AI-assisted comment moderation.

8 Analysis and Results

8.1 Gatekeeping by selectively opening comment sections

Analysis approach We first evaluated H1, RQ1a and RQ1b which concern gatekeeping by selectively opening up articles for commenting. We performed this analysis on the full dataset of 79,824 articles of which 19,090 (23.92%) had comment sections and 8850 (11.09%) were on social issues.

To test H1, we conducted a mixed effects logistic regression, modeling whether an article had a comment section or not (dependent variable) with an indicator variable identifying whether an article was on a social issue or not (independent variable). We controlled for the log-transformed page number where the article was published in the print edition (as a proxy for newsworthiness), if the article was published on a weekend (as news outlets have fewer moderation resources on weekends) and log-transformed article word length. We included random intercepts for the month during which the article was published and for the author of the article to account for our observation that certain months and certain authors are more likely to have articles with comment sections than others.

To test RQ1a and RQ1b, we conducted an identical regression but included an interaction between the social issue indicator and an indicator variable identifying whether the article was published before or after the adoption of AI-assisted moderation (evaluating RQ1a). In addition, we also included a random slope to account for variation in the effect of AI introduction across news sections (evaluating RQ1b). This random slope allows the association between the likelihood of having a comment section and the effect of AI moderation to vary across different news sections.

	Comment section present	
	Model 1	Model 2
Social issue (vs other)	0.230^{***}	0.336^{***}
	(0.034)	(0.059)
After AI (vs before AI)		1.228^{*}
		(0.489)
Social issue x After AI		-0.225^{**}
		(0.073)
Control variables		
Page number (log)	-0.502^{***}	-0.607^{***}
	(0.015)	(0.017)
Weekend	-0.884^{***}	-0.888^{***}
	(0.034)	(0.035)
Article length (log)	1.169^{***}	1.430^{***}
	(0.022)	(0.025)
Constant	-8.322^{***}	-10.246^{***}
	(0.181)	(0.503)
Observations	79,824	79,824
AIC	64,687	57,262
Marginal/Cond. R^2	0.168/0.593	0.205/0.737

Note: *p<0.05; **p<0.01; ***p<0.001

Table 1: Regression coefficients of models for evaluating H1, RQ1a and RQ1b

Results Table 1 shows the coefficients of the mixed effects binomial regression modeling the probability of an article having a comments section with (right column) and without (left column) the interaction between the after-AI-assisted moderation indicator and social issue article indicator. Reviewing the regression coefficient of the social issue article indicator in the left column (Model 1), we find that its association with the dependent variable is positive and significant ($\beta = 0.230, SE = 0.034$, $p\text{-}value < 0.001$). We find that the odds of a social issue article having a comments section is about 25.71% higher compared to other articles. This is contrary to our expectation that social issue articles are less likely to have comment sections. **H1 is not supported.**

Answering RQ1a, the right column (Model 2) in Table 1 shows the regression coefficients from the mixed effects logistic regression containing the interaction term. The regression coefficient of the interaction term is negative and statistically significant ($\beta = -0.225, SE = 0.073$, $p\text{-}value < 0.01$). We also include an interaction plot showing

[7] https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages?language=en_US

[8] We did not validate the likely to reject classifier as it aims to replicate NYT's moderation decisions which we are not privy to.

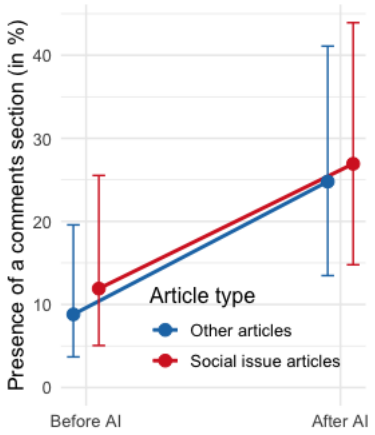

Figure 1: Plot of the interaction between the after ai indicator and the article type indicator. Note: The non-social issue articles experience a larger increase post AI compared to social issue articles.

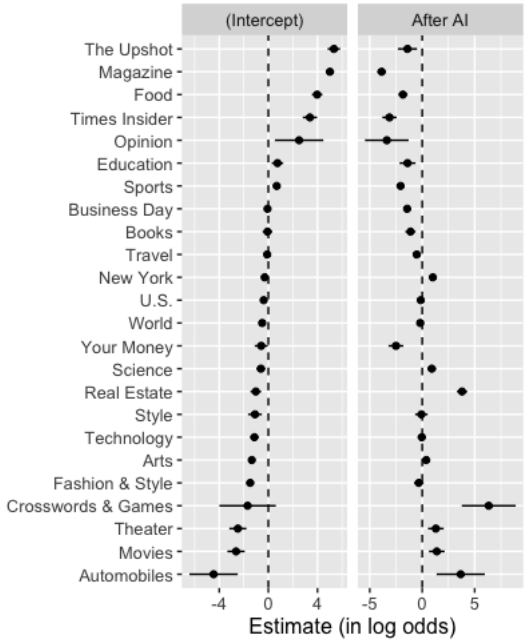

Figure 2: Plot of random slopes and intercepts from the mixed effects logistic regression modeling the probability of an article having a comments section.

the estimated marginal means in Figure 1. Based on posthoc comparisons, on average, 11.91% of social issue articles featured comment sections prior to the adoption of AI-assisted moderation, whereas 26.93% included comment sections thereafter ($OR = 2.73, SE = 1.34$, $p\text{-}value < 0.05$). In comparison, 8.81 % of non-social issue articles contained comment sections before AI adoption while 24.79% of non-social issue articles contained comment sections thereafter ($OR = 3.41, SE = 1.67$, $p\text{-}value < 0.05$), indicating a larger increase in the availability of comment sections for non-social issue articles than for social issue articles ($OR = 1.25, SE = 0.092$, $p\text{-}value < 0.01$).

Answering RQ1b, Figure 2 plots the random intercepts (left column) and slopes (right column) for different news sections (in log odds scale) from the mixed effects logistic regression (Model 2). The left column displaying the random intercepts for each section indicates the relative likelihood of having a comments section (compared to average) in the pre-AI period. The right column displaying the random slopes indicates the relative change in likelihood of having a comments section (compared to average change) in the post-AI period. Note that to enhance readability, we only include sections where either the intercept or slopes have p-value $<$ 0.05. In total, there are 34 sections of which 23 are shown.

Reviewing the left column, we find that prior to AI-assisted moderation introduction, many sections such as *Automobiles*, *Movies* and *Theater* with the least proportion of articles having comment sections are what can be called soft news. Interestingly, some sections that have a higher proportion of comment sections such as *Magazine* and *Food* can also be considered soft news. Hard news sections such as *U.S* or *World* where most political US and international news features in NYT appear to have comment sections slightly less often than average.

Reviewing the right column, we find that sections least likely to have comment sections before AI-assisted moderation appear to have gained the most. The sections with the largest gains include *Crosswords & Games*, *Real Estate*, *Automobiles*, *Movies* and *Theater*. Understandably, many of the sections that gained the least are the sections that already had a high proportion of articles with comment sections such as *Magazine* and *Opinion*. The correlation coefficient between the random intercept and the slope is -0.62, indicating a sizable negative correlation. The hard news sections such as *U.S.* and *World* recorded largely average gains.

Overall, AI-assisted comment moderation was associated with an increase in the availability of comment sections in soft news sections that previously had fewer comments sections. While the proportion of articles in hard news sections available for commenting also increased, this increase was not as drastic as in some soft news sections.

8.2 Gatekeeping through comment moderation

Analysis approach We evaluated the quality of the comments published before and after the adoption of AI-assisted comment moderation for only comments on social issue articles as we expected more potentially contentious and norm-violating comments on these articles. We sampled 5,000 comments from all those posted on social issue articles published in each month of our analysis period. In total, this resulted in 120,000 comments from 24 months. We obtained the probabilities for each quality measure using the Perspective API for these comments.[9] For classification purposes,

[9]We excluded a small number of comments (0.06%) for which the API did not return a response, likely because it did not support

	Attack on author	Attack on commenter	Incoherent	Inflammatory	Obscene	Spam	Unsubstantial	Toxicity
Before AI	0.91	0.79	0.77	0.71	0.96	0.99	0.73	0.87
After AI	0.94	0.80	0.81	0.73	0.97	0.99	0.76	0.87

Table 2: Accuracy of comment quality classifiers on comments published before and after AI-assisted moderation.

	Likely to reject	Attack on author	Attack on commenter	Incoherent	Inflammatory	Obscene	Spam	Unsubstantial	Toxicity
Before AI	27.34%	7.00%	19.81%	18.69%	30.54%	2.96%	2.45%	25.99%	1.89%
After AI	26.92%	7.06%	19.99%	19.86%	29.40%	2.86%	1.93%	27.14%	2.87%
z-test stat.	1.64	-0.44	-0.77	-5.11	4.34	1.02	6.23	-4.47	-11.16
p-value	0.101	0.656	0.439	< 0.001	< 0.001	0.307	< 0.001	< 0.001	< 0.001

Table 3: Proportion of comments by quality before and after AI-assisted moderation (two-sample z-test for proportions).

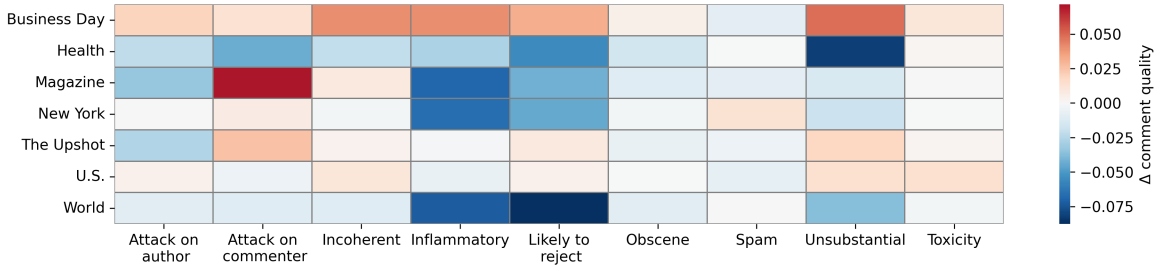

Figure 3: Heatmap showing the differences in comment quality by news section. Blue indicates improvement and red indicates a regression in comment quality.

we applied a probability threshold of 0.5.[10] To evaluate RQ2a, for each quality measure, we performed a two-sample z-test for proportions to compare the proportion of comments with that quality before and after AI-assisted moderation based on the classifier probability. To evaluate RQ2b, we examined the differences in comment quality by section. We restrict our analysis to sections that have at least 500 comments before and after AI assistance.

Results Table 3 shows the results of the z-tests. Although the differences across the time periods were found to be statistically significant for a few quality measures, in practical terms, these differences are rather small and are unlikely to have a meaningful impact (the largest difference across the measures between the two groups was 1.17%). Thus, overall, we observe that the quality of the published comments did not significantly change after the adoption of AI-assisted comment moderation.

However, Figure 3 reveals substantial heterogeneity in how comment quality changed across sections. In the heatmap, colors indicate the magnitude and direction of quality differences (After AI - Before AI), with red denoting declines in quality and blue indicating improvements following AI adoption. The *Business Day* section, which covers business and technology news, exhibits lower comment quality across nearly all dimensions. The *Magazine* section which hosts long-form journalism experienced an increase in comments attacking other commenters, alongside a modest decrease in comments targeting article authors. With the exception of *Business Day*, most sections show a marked decline or minimal change in inflammatory and likely-to-be-rejected comments. Finally, the two major hard news sections largely preserved (*U.S.*) or improved (*World*) comment quality after the introduction of AI assistance.

8.3 Changes in audience engagement in the comment sections

Analysis approach To evaluate RQ3a, we compare the average number of articles commented on by users per month before and after AI-assisted moderation. We conducted a random effects negative binomial regression modeling the number of unique articles commented on by a user per month for the analysis time period. We included a random effect for the month when the user commented and added an indicator for whether the comment was posted before or after the adoption of AI-assisted comment moderation (independent variable). We performed this comparison only among users who have commented on at least one article in that month as we do not have information on users who did not comment on an article. To evaluate RQ3b, we conducted an identical regression and included an interaction term between the after AI moderation indicator and commenter subgroup indicator. For each month during our analyzing period, users were classified into three

the language detected in the comments.

[10]Varying this [0.5, 0.7] does not substantively change results.

	Num. of articles commented	
	Model 1	Model 2
After AI (vs before AI)	0.157***	0.063
	(0.024)	(0.036)
User type (vs other users)		
Least active		-1.639***
		(0.007)
Most active		1.631***
		(0.006)
Interaction terms		
After AI x Least active		-0.032***
		(0.008)
After AI x Most active		0.179***
		(0.007)
Constant	0.301	0.482***
	(0.017)	(0.026)
Observations	1,012,292	1,012,292
AIC	2,919,898	2,468,896
Marginal/Cond. R2	0.009/0.015	0.606/0.610

Note: *p<0.05; **p<0.01; ***p<0.001

Table 4: Regression coefficients of the regression for RQ3a and RQ3b

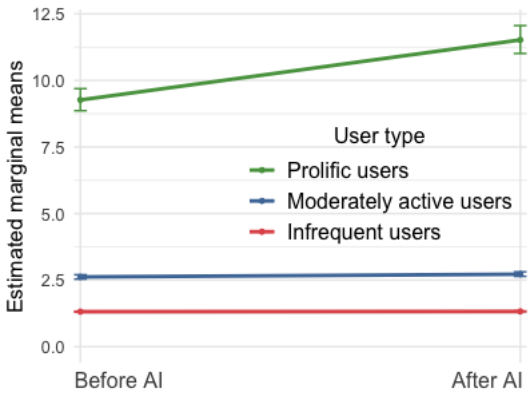

Figure 4: Plot of the interaction between the after AI indicator and user type modeling the average number of articles commented on.

groups based on the number of unique articles they commented on over the previous six months: (i) prolific commenters: those whose commenting frequency exceeded the mean by more than two standard deviations (average monthwise $SD = 13.20$) (ii) infrequent commenters: users who commented on two articles or fewer[11], (iii) moderately active commenters which included all other users. Finally, to answer RQ3c, we compare different commenter groups across the two time periods using two-sample z-test for proportions.

Results Table 4 shows the coefficients of the random effects negative binomial regression modeling the number of articles commented on by each user per month with (right column) and without (left column) the interaction between the after AI indicator and user group type indicator indicator. Answering RQ3a, from the left column, the regression coefficient for the after AI indicator is positive and statistically significant ($\beta = 0.157, SE = 0.024, p\text{-}value < 0.01$). This suggests that the adoption of AI-assisted moderation was associated with a 16.97% increase in the expected number of articles commented on per user per month. The right column in Table 4 shows the coefficients of interaction terms to be statistically significant and Figure 4 shows the corresponding interaction plot. Conducting planned contrasts to answer RQ3b, we found that prolific commenters, on average, commented on 11.52 articles per month post-AI compared to 9.27 articles pre-AI comment moderation. This difference was statistically significant ($OR = 1.27, SE = 0.046, z\text{-}ratio = 6.701, p\text{-}value < 0.001$). Differences in the average number of articles commented on by the other two user groups during the pre and post AI-assisted comment moderation periods were not statistically significant.

RQ3c examines the proportion of user subgroups before and after the introduction of AI-assisted moderation. A comparison of user composition before and AI introduction shows that prolific, infrequent, and moderately active commenters accounted for 6.71% vs. 6.92% ($z = -3.95$, $p\text{-}value < 0.001$)[12], 58.86% vs. 51.90% ($z = -67.49$, $p\text{-}value < 0.001$), and 34.43% vs. 41.18% ($z = 68.01$, $p\text{-}value < 0.001$) of commenters, respectively. This pattern indicates a compositional shift in participation across subgroups. Post AI, while the share of prolific commenters remained largely unchanged, the proportion of infrequent commenters decreased alongside an increase in the proportion of moderately active commenters. We discuss the implications of these results in the discussion section.

9 Discussion

The results of our analyses shed new light on shifting media gatekeeping logic and audience engagement following the transition to AI-assisted comment moderation. First, contrary to our hypothesis, we found that social issue articles include comment sections more often than other articles. This challenges the conventional wisdom that media organizations generally restrict comments on social issues to avoid potentially uncivil discussions (Santana 2016). We propose two explanations for why the organization may diverge from the broader industry trend of limiting comment sections on social issue articles. First, NYT is among the most well-resourced news organizations globally, which likely allows them to allocate substantial moderation resources to handle potentially hostile content. Second, NYT employs a pre-moderation system for its comments. This means that all audience-generated content is reviewed before being published, ensuring that uncivil content does not appear on the

[11]The median articles commented on in the prior six months is 2. Varying this threshold [1, 3] does not change the results.

[12]Though statistically significant, the practical significance of this difference is likely minor.

site. As a result, there is little risk to its brand reputation from opening up comment sections on social issues.

We found that the adoption of AI-assisted comment moderation was associated with a disproportionate increase in comment sections in soft news sections such as *Real Estate* and *Automobiles* compared to hard news section such as *U.S.* and *World* section which contain more political news. This mirrors our finding that the adoption of AI had a more substantial effect on the availability of comment sections in non-social issue articles compared to social issue articles. These results suggest that AI-assisted comment moderation was associated with opening up more articles for commenting that are relatively lighter in tone and arguably, substance. For example, *Crosswords & Games*, an area that NYT has heavily invested in over the past decade (Maher 2023), recorded the highest gains in comment sections among all news sections. Commenting opportunities in such soft news spaces where there is relatively little conflict may foster a sense of community among readers that, among other benefits, may help attract and retain subscribers. In contrast, AI adoption was linked to a relatively limited increase in the number of social issue articles opened for public discussion. Given evidence linking online political discussions to normatively positive democratic outcomes, including increased exposure to alternate views (Stromer-Galley 2003) and higher opinion quality (Price, Nir, and Cappella 2006), this pattern is concerning. We posit that social issue content may remain relatively resource-intensive to moderate, even when supported by AI, as comments on these articles require rigorous fact-checking to guard against misinformation (Udupa, Maronikolakis, and Wisiorek 2023) and often address domains such as race and gender, where current AI systems demonstrate well-documented limitations (Nakka 2025). In contrast, moderation decisions in soft news topics may be more reliably evaluated by AI, likely contributing to the differences we observe in opening up more articles for commenting in social and non-social issue domains.

Comparing comment quality before and after the introduction of AI-assisted moderation, we find little change in aggregate outcomes; however, substantial heterogeneity emerges across news sections, potentially reflecting section-specific moderation practices (McInnis et al. 2021). It is important to note that we observe news-section wise differences in spite of evaluating only social issue articles across all sections. In especially contentious sections such as *World*, which features global political content, the prevalence of inflammatory and likely-to-be-rejected comments either declined significantly. This pattern could reflect a reduction in the submission of problematic content in the post–AI period, although this explanation appears unlikely given evidence of increasing polarization during the analysis period (Boxell, Gentzkow, and Shapiro 2024). A more plausible explanation is that these patterns are a result of more heightened moderation scrutiny in these sections. However, our analysis cannot conclusively determine whether this scrutiny is directly attributable to AI-assisted moderation or instead to concurrent editorial or policy decisions to reduce problematic content in these sections.

Finally, in terms of audience engagement, we find that the increased availability of comment sections is associated with a measurable increase in the average number of articles on which users choose to comment. However, subgroup analysis suggests that users who were already prolific commenters significantly increased their participation while other less active commenters did not exhibit meaningful changes in their engagement levels following the introduction of AI-assisted moderation. We speculate that by reducing moderation bottlenecks and expanding the availability of comment sections, AI introduction may have allowed the most engaged users to participate more. While we do not observe an increase in the proportion of prolific commenters, we do observe a significant decrease in proportion of infrequent commenters and a corresponding increase in moderately active commenters. This pattern may be interpreted in multiple ways and more research is needed to disentangle whether it reflects, for example, a conversion of previously infrequent commenters to moderately active commenters or simply a disproportionate reduction in infrequent commenters. Regardless of the underlying mechanisms, shifts in the composition of participating commenters may influence the nature of discussions, as users with different activity levels tend to contribute in distinct ways (Graham and Wright 2014). For instance, political communication research indicates that highly vocal social media users are disproportionately partisan (Krupnikov and Ryan 2022). Overall, these results suggest that the introduction of AI assistance was associated with significant changes in engagement patterns; not only in the volume of participation but also in how participation is distributed across different segments of the commenting audience.

10 Limitations

We acknowledge the limitations of this study. While the NYT's early adoption of a hybrid comment moderation system and reputation for high-quality comments made it an ideal case study, its distinctive characteristics in terms of its sizable resources and pre-moderation practices may limit the generalizability of our findings.

The analysis relies on real-world observational data rather than a controlled experiment. Although this strengthens external validity, it does not support causal inference and leaves open the possibility of unobserved confounding. The study period coincided with major external developments such as Supreme Court nominations and immigration enforcement debates, as well as significant changes within NYT, including rapid subscriber growth. These dynamics may have contributed to differences observed between the pre- and post-AI moderation periods. While election timelines are incorporated into the study design, residual confounding may remain. The absence of a viable control group further limits causal interpretation. Future work comparing the NYT comments section to other platforms that did not transition to using AI assistance for moderation can help isolate the true effect of AI assistance. The results should therefore be viewed as suggestive and correlational evidence rather than conclusive and causal.

Changes within NYT may affect some of the presented results. While the NYT typically closes comments sections 24 hours after opening, in some instances, moderators may

close sections if they sense that "there is nothing substantial to gain from more comments on the article."[13] These differences in when a comment section is closed may affect our audience metrics analyses especially if the frequency of such early closures varied before and after AI adoption. Similarly, changes to NYT's article-to-keyword mapping may adversely affect our ability to identify social issue articles.

Ideally, for gatekeeping analyses, we would have access to all inputs submitted to the gatekeeper (eg. all submitted comments) and all outputs after gatekeeping (eg. published comments after moderation decisions). We have this level of visibility for our analysis of gatekeeping via selective comment section openings. However, for comment moderation analysis, we can only observe the output, that is, comments that were approved and published. This prevents us from evaluating whether the quality of submitted comments or the moderation criteria have shifted over time. Nonetheless, by analyzing the output, we can still draw meaningful inferences about the change in the quality of published comments before and after the adoption of AI-assisted moderation.

Finally, although we validated the Perspective API across eight dimensions and two time periods, it is possible that the classifiers have some inherent biased in their outputs (Udupa, Maronikolakis, and Wisiorek 2023) which may affect the results of our analyses if the bias was unevenly distributed across the two time periods. Further, we did not account for more granular news-section specific differences in classifier accuracies which may affect the results for RQ2b.

11 Conclusion

Overall, the adoption of AI-assisted comment moderation at NYT was associated with shifts in gatekeeping and audience engagement. Commenting opportunities expanded unevenly across news domains, while overall comment quality remained stable despite section-level differences. In terms of audience engagement, these results suggest that AI-assisted moderation coincided with changes not only in the level of participation, but also in how participation was distributed across users, subtly shaping public discourse.

Acknowledgements

We thank Ayman Mahfuz, Mayah Piunno and Tanvi Prem for research assistance. We thank Moo Sun Kim, Shuting Yao, Gina Masullo, Jo Lukito, Craig Scott, Shengchun Huang and others at the Center for Media Engagement for their thoughtful suggestions on prior versions of this paper. We also thank the anonymous reviewers for their constructive feedback during the review process. The authors also thank the John S. and James L. Knight Foundation for funding that supported this work.

References

Bächtiger, A.; Dryzek, J. S.; Mansbridge, J.; and Warren, M. 2018. Deliberative democracy. *The Oxford handbook of deliberative democracy*, 1–32.

Boberg, S.; Schatto-Eckrodt, T.; Frischlich, L.; and Quandt, T. 2018. The moral gatekeeper? Moderation and deletion of user-generated content in a leading news forum. *Media and Communication*, 6(4): 58–69.

Boxell, L.; Gentzkow, M.; and Shapiro, J. M. 2024. Cross-country trends in affective polarization. *Review of Economics and Statistics*, 106(2): 557–565.

Bruns, A. 2021. Gatewatching and news curation. In *The Routledge Companion to Political Journalism*, 252–261.

Center, P. R. 2016. 4. Top voting issues in 2016 election.

Center, P. R. 2020. 4. Important issues in the 2020 election.

Diakopoulos, N. A. 2015. The editor's eye: Curation and comment relevance on the New York times. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 1153–1157.

Engelke, K. M. 2020. Enriching the conversation: Audience perspectives on the deliberative nature and potential of user comments for news media. *Digital journalism*, 8: 447–466.

Etim, B. 2017. The Times Sharply Increases Articles Open for Comments, Using Google's Technology.

Gilbert, E. 2013. Widespread underprovision on reddit. In *Proceedings of the 2013 conference on Computer supported cooperative work*, 803–808.

Graham, T.; and Wright, S. 2014. Discursive equality and everyday talk online: The impact of "superparticipants". *Journal of Computer-Mediated Communication*, 19.

He, L.; Han, C.; Mukherjee, A.; Obradovic, Z.; and Dragut, E. 2020. On the dynamics of user engagement in news comment media. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(1): e1342.

He, Y.; Gu, Y.; Shekhar, R.; Castro, I.; and Tyson, G. 2024. Making the Pick: Understanding Professional Editor Comment Curation in Online News. In *International AAAI Conference on Web and Social Media*, volume 18.

Hermida, A.; and Thurman, N. 2008. A clash of cultures: The integration of user-generated content within professional journalistic frameworks at British newspaper websites. *Journalism practice*, 2(3): 343–356.

Horta Ribeiro, M.; Cheng, J.; and West, R. 2023. Automated content moderation increases adherence to community guidelines. In *Proceedings of the ACM web conference 2023*, 2666–2676.

Ihlebæk, K. A.; and Krumsvik, A. H. 2015. Editorial power and public participation in online newspapers. *Journalism*.

Jensen, E. 2016. NPR Website To Get Rid Of Comments.

Juarez Miro, C. 2022. The comment gap: Affective publics and gatekeeping in The New York Times' comment sections. *Journalism*, 23(4): 858–874.

Koelmann, H.; Müller, K.; Niemann, M.; and Riehle, D. M. 2022. Moderating the Good, the Bad, and the Hateful: Moderators' Attitudes Towards ML-based Comment Moderation Support Systems. In *Multidisciplinary International Symposium on Disinformation in Open Online Media*, 100–113.

Krupnikov, Y.; and Ryan, J. B. 2022. *The other divide*. Cambridge University Press.

[13] https://archive.nytimes.com/publiceditor.blogs.nytimes.com/2012/10/15/questions-and-answers-on-how-the-times-handles-online-comments-from-readers/

León, E. d.; Vermeer, S.; and Trilling, D. 2023. Electoral news sharing: a study of changes in news coverage and Facebook sharing behaviour during the 2018 Mexican elections. *Information, communication & society*, 26(6): 1193–1209.

Maher, B. 2023. How games are powering online subscriptions at The New York Times.

Masullo, G. M.; Tenenboim, O.; and Lu, S. 2023. "Toxic atmosphere effect": Uncivil online comments cue negative audience perceptions of news outlet credibility. *Journalism*, 24(1): 101–119.

McInnis, B.; Ajmani, L.; Sun, L.; Hou, Y.; Zeng, Z.; and Dow, S. P. 2021. Reporting the Community Beat: Practices for Moderating Online Discussion at a News Website. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–25.

Meltzer, K. 2015. Journalistic concern about uncivil political talk in digital news media: Responsibility, credibility, and academic influence. *The International Journal of Press/Politics*, 20(1): 85–107.

Mohammadi, S.; and Yasseri, T. 2025. AI Feedback Enhances Community-Based Content Moderation through Engagement with Counterarguments. *arXiv:2507.08110*.

Molina, M. D.; and Sundar, S. S. 2022. When AI moderates online content: effects of human collaboration and interactive transparency on user trust. *Journal of Computer-Mediated Communication*, 27(4): zmac010.

Muddiman, A.; and Stroud, N. J. 2017. News values, cognitive biases, and partisan incivility in comment sections. *Journal of communication*, 67(4): 586–609.

Nakka, N. 2025. An Evaluation of the Google Perspective API by Race and Gender. In *Proceedings of the 17th ACM Web Science Conference 2025*, 522–527.

Niemann, M. 2021. Elicitation of requirements for an AI-enhanced comment moderation support system for non-tech media companies. In *International Conference on Human-Computer Interaction*, 573–581. Springer.

O'Brien, S. A. 2017. The New York Times wants you to read the comments.

Paasch-Colberg, S.; and Strippel, C. 2022. "The boundaries are blurry…": how comment moderators in Germany see and respond to hate comments. *Journalism Studies*, 23(2).

Pierson, E. 2015. Outnumbered but well-spoken: Female commenters in the New York Times. In *Proceedings of the 18th acm conference on computer supported cooperative work & social computing*, 1201–1213.

Price, V.; Nir, L.; and Cappella, J. N. 2006. Normative and informational influences in online political discussions. *Communication Theory*, 16(1): 47–74.

Reinemann, C.; Stanyer, J.; Scherr, S.; and Legnante, G. 2012. Hard and soft news: A review of concepts, operationalizations and key findings. *Journalism*, 13(2): 221–239.

Rieder, B.; and Skop, Y. 2021. The fabrics of machine moderation: Studying the technical, normative, and organizational structure of Perspective API. *Big Data & Society*.

Ruckenstein, M.; and Turunen, L. L. M. 2020. Re-humanizing the platform: Content moderators and the logic of care. *New media & society*, 22(6): 1026–1042.

Salganik, M. J.; and Lee, R. C. 2020. To Apply Machine Learning Responsibly, We Use It in Moderation. The New York Times (Open).

Santana, A. D. 2016. Controlling the conversation: The availability of commenting forums in online newspapers. *Journalism studies*, 17(2): 141–158.

Shoemaker, P. J.; Eichholz, M.; Kim, E.; and Wrigley, B. 2001. Individual and routine forces in gatekeeping. *Journalism & mass communication quarterly*, 78(2): 233–246.

Shoemaker, P. J.; and Reese, S. D. 2013. *Mediating the message in the 21st century: A media sociology perspective*.

Shoemaker, P. J.; and Vos, T. 2009. *Gatekeeping theory*.

Soroka, S. N. 2012. The gatekeeping function: Distributions of information in media and the real world. *The Journal of Politics*, 74(2): 514–528.

Stromer-Galley, J. 2003. Diversity of political conversation on the Internet: Users' perspectives. *Journal of Computer-Mediated Communication*, 8(3): JCMC836.

Udupa, S.; Maronikolakis, A.; and Wisiorek, A. 2023. Ethical scaling for content moderation: Extreme speech and the (in) significance of artificial intelligence. *Big Data & Society*, 10(1): 20539517231172424.

Wang, S. 2023. Factors related to user perceptions of artificial intelligence (AI)-based content moderation on social media. *Computers in Human Behavior*, 149: 107971.

Wang, Y.; and Diakopoulos, N. 2022. Highlighting high-quality content as a moderation strategy: The role of new york times picks in comment quality and engagement. *ACM Transactions on Social Computing (TSC)*, 4(4): 1–24.

Waterschoot, C. 2024. 3. Governing the "Third Half of the Internet": The Dynamics of Human and AI-Assisted Content Moderation. *the Digital Society*, 63.

Watson, B. R.; Peng, Z.; and Lewis, S. C. 2019. Who will intervene to save news comments? Deviance and social control in communities of news commenters. *New media & society*, 21(8): 1840–1858.

Weisner, A.; Schäfer, S.; and Lecheler, S. 2025. Navigating the gray areas of content moderation: Professional moderators' perspectives on uncivil user comments and the role of (AI-based) technological tools. *new media & society*, 27(3).

Wong, K.; Paritosh, P.; and Aroyo, L. 2021. Cross-replication Reliability-An Empirical Approach to Interpreting Inter-rater Reliability. In *Proceedings of 59th Annual Meeting of the Association for Computational Linguistics*.

Yu, Z.; Otto, L.; Assenmacher, D.; and Wagner, C. 2024. A systematic review of the effects of ai-assisted moderation on individuals and groups. *Human-Machine Communication*.

Paper Checklist

1. For most authors...

(a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes. We address the potential risks to user privacy by only releasing aggregate anonymized comment data, including text classifier outcomes (without the underlying text) necessary for replicating our results.

(b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes

(c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes

(d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes. We note that commenters are unrepresentative of the general population.

(e) Did you describe the limitations of your work? Yes

(f) Did you discuss any potential negative societal impacts of your work? Yes, see our brief ethical considerations statement below.

(g) Did you discuss any potential misuse of your work? Yes, see our brief ethical considerations statement below.

(h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes

(i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes

2. Additionally, if your study involves hypotheses testing...

(a) Did you clearly state the assumptions underlying all theoretical results? Yes

(b) Have you provided justifications for all theoretical results? Yes

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes

(e) Did you address potential biases or limitations in your theoretical framework? Yes

(f) Have you related your theoretical results to the existing literature in social science? Yes

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? NA

(b) Did you include complete proofs of all theoretical results? NA

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, we intend to do so when accepted for publication.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA, we used existing classifiers through an API

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA, we used existing classifiers through an API

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA, we use an existing classifiers through an API

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, see our brief ethical considerations statement below.

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? NA

(b) Did you mention the license of the assets? NA

(c) Did you include any new assets in the supplemental material or as a URL? NA

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, see our brief ethical considerations statement below.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, see our brief ethical considerations statement below.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? We have discussed releasing the datasets according to FAR.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? We will be creating a Datasheet once the paper is accepted for publication.

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA

A Additional Ethics Statement

As with most computational research on large datasets, we are unable to obtain informed consent from users in our dataset. To mitigate potential risks to privacy, we intend to only share (i) complete article metadata indicating which articles feature comment sections (containing no user-generated content), and (ii) aggregate anonymized comment data, including text classifier outcomes (without the underlying text), necessary for replication. We do not share the usernames or comment texts as they may contain personally identifiable information. We believe our data handling approach minimizes risks to user privacy while still facilitating reproducibility.

A potential negative societal impact of this work might be that users self-censor and comment less often in the platform because of privacy concerns. We aimed to mitigate this issue by minimizing risks to privacy. Further, we believe this study may encourage more user participation by assuaging users about certain biases in AI-assisted comment moderation as we show that the comment publishing standards have not significantly changed. A potential misuse of the work could stem from mining user comments. However, by only releasing aggregate comment data, we are able to address these concerns.

Finally, we discuss the costs of misclassification. In this study, misclassifications may arise in two cases: (i) when we incorrectly classify an article as a social issue or not. (ii) when we label the quality of comment based on the eight quality measures incorrectly. Misclassifying articles may skew the analysis on the availability of comment sections. We aimed to mitigate this issue by manually evaluating the classifications and ensuring that they are of reasonable accuracy. The errors in comment quality analysis may lead to biased assessments of AI-assisted comment moderation. While we did not formally evaluate these eight measures, we performed robustness checks to ensure that varying the thresholds do not substantively change the results.

B Months Selected for Analyses

Pre-AI months	Post-AI months
2013-08-01	2017-08-01
2013-10-01	2017-10-01
2013-11-01	2017-11-01
2014-07-01	2018-07-01
2014-09-01	2018-09-01
2014-12-01	2018-12-01
2015-02-01	2019-02-01
2015-04-01	2019-04-01
2015-06-01	2019-06-01
2015-10-01	2019-10-01
2015-12-01	2019-12-01
2016-04-01	2020-04-01

Table 5: Randomly selected pre- and post-AI adoption months

C Prompt Used to Identify Social Issue Articles

You are an AI assistant. You are provided with an article title, lead paragraph, and abstract. Your task is to determine whether the article is about CANDIDATE_TOPIC or not. Analyze the article's content including the title, lead paragraph, and abstract to identify key themes or references to CANDIDATE_TOPIC. Use logical reasoning to confirm the presence or absence of relevant content. Conclude by classifying the article and providing your reasoning.
Steps:

1. Review the title, lead paragraph, and abstract to extract context and key themes.
2. Look for keywords or phrases typically related to the topic of CANDIDATE_TOPIC.
3. Assess whether the identified keywords or themes explicitly refer to or imply discussions on CANDIDATE_TOPIC.
4. Decide whether the article discusses CANDIDATE_TOPIC based on your analysis.
5. Record the reasoning process that led you to your conclusion.

Output Format: Return the result in the following JSON format: {class: Yes or No, reason: your detailed reasoning here.}
Notes:
- Address any ambiguity by explaining the rationale behind the classification decision.
- Ensure clarity in your reasoning to substantiate your choice between 'Yes' or 'No'."
User input: Headline: TITLE, Lead paragraph: LEAD_PARAGRAPH Abstract: ABSTRACT

Table 6: Social issues and corresponding keywords

Issue	**Keywords**
Abortion	Abortion
Climate change	Carbon Caps and Emissions Trading Programs; Clean Air Act; Coast Erosion; Earth Day; Global Warming; Wind Power
Economic inequality	Food Stamps; Income Inequality; Living Wage; Minimum Wage; Poverty; Welfare (US)
2nd amendment	Gun Control
Homelessness	Homeless Persons
Immigration	Asylum, Right of; Citizenship and Naturalization; Deferred Action for Childhood Arrivals; Deportation; Family Separation Policy (US Immigration); Foreign Students (in US); Foreign Workers; Illegal Immigration; Immigration Detention; Immigration and Emigration; Middle East and Africa Migrant Crisis; Refugees and Displaced Persons
LGBTQ issues[13]	Homosexuality; Homosexuality and Bisexuality; Same-Sex Marriage, Civil Unions and Domestic Partnerships; Stonewall Riots (1969); Transgender and Transsexuals
Universal healthcare	Health Insurance and Managed Care
Police Brutality	Police Brutality, Misconduct and Shootings
Race issues	Affirmative action; Anti-semitism; Apartheid (Policy); Asian-Americans; Black People; Charlottesville, Va, Violence (August, 2017); Civil Rights Movement (1954-68); Hate Crimes; Indigenous Australians; Minorities; Muslim Americans; Native Americans; Race and Ethnicity; Segregation and Desegregation; Slavery; Slavery (Historical); United States National Anthem Protests (2016-); Whites
Sexual misconduct	#MeToo Movement; Sex Crimes; Sexual Harassment

[13] We note that LGBTQ+ communities do not necessarily use these terms to describe themselves and that some of the terms may be considered offensive. These labels were drawn from The New York Times' keyword classification system and do not reflect the views or language preferences of the authors.