Media and Communication
2026 • Volume 14 • Article 12358
https://doi.org/10.17645/mac.12358

cogitatio

ARTICLE

Open Access Journal

Hybrid Moderation in Practice: How Users, AI, and Professionals Judge Uncivil Comments

Aviv Barnoy

Department of Media and Communication, Erasmus University Rotterdam, The Netherlands

Correspondence: Aviv Barnoy (barnoy@eshcc.eur.nl)

Submitted: 10 March 2026 **Accepted:** 25 June 2026 **Published:** 30 July 2026

Issue: This article is part of the issue "The Role of AI for Counter Speech: Detection, Intervention, and Risks" edited by Lena Frischlich (University of Southern Denmark – Odense), Cathy Buerger (Dangerous Speech Project), and Magdalena Obermaier (LMU Munich), fully open access at https://doi.org/10.17645/mac.i500

Abstract

Online platforms increasingly rely on hybrid moderation systems that combine automated triage with professional human review, yet little research compares how these actors judge the same real content in production-linked settings. This article examines uncivil online comments as a platform-governance problem by comparing three decision-makers: registered users, a platform AI system, and professional moderators. Using a secondary dataset from a US-based publisher comment environment, we analyze real comments spanning five moderation categories: Violence, Doxxing, Hate Speech, Sexual Explicitness, and Personal Attacks. The dataset combines production-stage AI workflow dispositions, professional moderators' final adjudications, and ex-post survey judgments from 863 registered users, yielding 8,588 user-comment decisions. Results show that users chose to allow publication of roughly two-thirds of the uncivil comments and that their decisions aligned more closely with the AI system's publish-and-moderate versus hold-and-moderate orientation than with professional moderators' publish-versus-block decisions. Divergence was concentrated in specific categories and items, especially Hate Speech. Users' justifications suggest that permissive decisions often relied on threshold reasoning, rejection of toxicity, or free-speech concerns, whereas restrictive decisions more often relied on direct violation labels. Demographic and engagement-based predictors explained little variance in alignment. The findings suggest that users, AI systems, and professional moderators should not be treated as interchangeable moderation tools: they operationalize different thresholds and may be suited to different governance functions.

Keywords

AI-assisted moderation; hybrid moderation; platform governance; uncivil comments; uncivil online communication; user-judgments

1. Introduction

Harmful and uncivil online communication continues to pose a challenge for public discourse. Platforms that host large-scale conversation are expected to curb harmful speech, yet doing so at scale is difficult: Enforcement cannot depend solely on user reporting or manual review, and the volume of user-generated content pushes toward at least some automation (Arora et al., 2024; Matamoros-Fernández et al., 2023; Suzor et al., 2018). Furthermore, not every rude, harsh comment raises the same democratic or safety concerns. Prior work has repeatedly cautioned against treating incivility, toxicity, intolerance, and harmfulness as interchangeable concepts (Bormann et al., 2022; Kümpel & Unkel, 2023; Rossini, 2020, 2022). Different problematic comments may call for different thresholds, decision-makers, and contextual sensitivity.

These challenges are not only technical. Content moderation is increasingly treated as a problem of platform governance: Platforms exercise private power over speech, and their policies and practices can be assessed against legitimacy-relevant values such as transparency, accountability, due process, and respect for rights (Katsaros et al., 2024; Suzor et al., 2018).

Moderation is also structurally trade-off laden. Choices routinely involve competing goals and values, such as harm reduction versus participation, efficiency versus quality, and consistency versus contextual discretion. These trade-offs include explicit tensions between human and automated approaches and between centralized and distributed governance arrangements (Jiang et al., 2023). Thus, moderation decisions are not merely classifications of content; they are practical judgments about which risks, norms, and values should be prioritized in a given communicative environment.

In practice, many platforms rely on hybrid regimes that distribute authority across different decision-makers: algorithmic systems, professional moderators, and participatory mechanisms that incorporate users (Barnoy, Freiman, et al., 2025; Yu et al., 2024). Empirical work on "wisdom of the crowds" in moderation contexts illustrates the centrality of crowd inputs for surfacing potential problems, while also documenting their limitations and vulnerability to misuse (Bozarth et al., 2023). Platform studies of automated moderation tools in large communities similarly suggest that algorithmic systems can reshape human moderation practice rather than simply replacing it (He et al., 2025). These actors should therefore not be understood as interchangeable routes to the same moderation outcome. They may represent different moderation logics.

Although the present study does not examine counterspeech in the narrow sense of producing a direct reply, it examines an upstream governance stage in which users, AI systems, and professional moderators shape which comments are removed, left visible, or potentially made available for counter-response. Thus, a key empirical question is not only whether moderation is effective, but whose judgment is effectively instantiated when authority is distributed across users, automated systems, and professional moderators. Alignment is important in this context not because agreement is inherently desirable, or because one actor should automatically be treated as a gold standard, but because patterns of convergence and divergence reveal which standards are being enacted in practice. If users, AI systems, and professional moderators operationalize harmfulness differently, then hybrid moderation systems need to be designed around these differences rather than assuming that one actor can straightforwardly substitute for another. Some websites, communities, or content types may require stronger professional adjudication; others may benefit from community-sensitive thresholds; and others may depend on AI primarily as a scalable triage mechanism.

Yet much of the research on AI and moderation is forced to study proxies for real governance: hypothetical scenarios (e.g., Kozyreva et al., 2023), synthetic examples, technical benchmarks, or user perceptions detached from production workflows (Arora et al., 2024; Yu et al., 2024). Even where legitimacy is studied directly, the focus is often on audiences' perceptions (e.g., Molina & Sundar, 2022; Suzor et al., 2018), rather than on how different decision-makers actually converge or diverge on the same real content. This leaves a gap in our understanding of what hybrid moderation systems do in practice: When automated triage and professional adjudication operate together, are the resulting judgments closer to professional standards, closer to community thresholds, or systematically different from both?

To address this gap, we draw on a norm-based approach to problematic online discourse. In this view, terms such as "toxicity" are useful only when connected to specific norm violations rather than treated as broad evaluative labels (Keren et al., 2025). The broader framework distinguishes between different foundations of healthy online conversation, including civil and epistemic norms. The present article, however, focuses only on uncivil comments that raise civil, interpersonal, or safety-related moderation concerns. Specifically, we examine five platform-relevant categories: Violence, Doxxing, Hate Speech, Sexual Explicitness, and Personal Attacks. We therefore use the norm-violation framework not to collapse these categories into a single concept of "toxicity," but to ask how different moderation actors enforce, tolerate, or contest different forms of potentially harmful or uncivil content.

Empirically, the article leverages rare access to a production-linked moderation dataset connecting (a) automated moderation, (b) professional moderators' decisions, and (c) judgments from a large sample of registered users evaluating the same real comments. Our goals are to: (a) characterize baseline community thresholds for publication versus removal across key uncivil categories; (b) compare how user judgments align with automated moderation and professional adjudication; and (c) map where convergence and divergence concentrate, including user-level heterogeneity and users' stated justifications.

2. Theoretical Framework

2.1. Uncivil Comments as Norm Violations

"Toxicity" in online discourse is often discussed through broad evaluative language, such as "unhealthy," "low-quality," or "harmful" discourse. Such language can be useful, but it can also obscure important conceptual distinctions. In this article, we adopt Keren et al.'s (2025) approach, viewing uncivil comments through the lens of norm violation: Comments become moderation-relevant when they are understood as breaching acceptable standards, such as safety, dignity, respect, and epistemic responsibility (Barnoy, Freiman, et al., 2025; Bormann et al., 2022; Keren et al., 2025). This approach does not imply that all norm violations are equivalent or that all uncivil content should be removed. Rather, it provides a way to connect platform categories, user judgments, and moderation decisions to the norms they appear to enforce or tolerate.

This distinction is also important because prior research has repeatedly shown that incivility, intolerance, and toxicity are often used inconsistently. Some studies treat incivility primarily as rudeness, profanity, or interpersonal disrespect; others include more severe violations of democratic or moral norms, such as hate speech, threats, or discriminatory stereotyping (Bormann et al., 2022; Kümpel & Unkel, 2023; Rossini, 2020,

2022). Rossini (2020, 2022) argues for distinguishing incivility, understood largely as rude or harsh discourse, from intolerance, which targets individuals or groups in ways that violate moral respect and democratic pluralism. Kümpel and Unkel (2023) show that users react more negatively to intolerant comments than to merely uncivil ones. Masullo Chen et al. (2019) caution that treating incivility as something that should always be eliminated risks ignoring the contextual and sometimes democratically useful role of heated or norm-challenging speech.

The five categories examined in this study should therefore not be read as equivalent subtypes of "incivility." Personal Attacks are closest to interpersonal incivility; Hate Speech and Violence are closer to intolerance or safety-related harms; Doxxing concerns privacy and invasive exposure; and Sexual Explicitness often reflects platform-specific standards of appropriateness, audience protection, and safety. We treat these categories as uncivil moderation categories: platform-relevant categories of potentially problematic content that raise civil, interpersonal, or safety-related concerns, while recognizing that they differ in severity, normative basis, and likely enforcement threshold. Throughout the remainder of the article, we use "uncivil" as a shorthand for these five moderation categories, not as a claim that all five are equivalent instances of incivility in the narrow theoretical sense.

This framing motivates our first research question (RQ), which establishes baseline community judgments about uncivil comments before turning to comparisons with AI and professional moderation:

> RQ1: How often do users decide to publish vs. block uncivil comments across the five focal moderation categories (Violence, Doxxing, Hate Speech, Sexual Explicitness, Personal Attacks)?

2.2. Alignment as a Governance Question in AI-Assisted Moderation

To manage uncivil speech at scale, online platforms increasingly rely on hybrid moderation regimes that combine professional moderators, algorithmic systems, and forms of community participation. As a result, the central governance question is no longer only what content should be restricted, but also whose judgment is effectively implemented when decisions are distributed across humans and machines. This shift is consequential because content moderation is widely recognized as value-laden decision-making rather than a purely technical filtering task. Jiang et al.'s (2023) trade-off-centered synthesis conceptualizes moderation as a set of recurring trade-offs at multiple levels, including actions, styles, philosophies, and values. These trade-offs include explicit tensions between human and automated moderation, centralized and distributed governance, and efficiency and quality. In practice, these choices determine whether a system prioritizes consistency, speed, contextual discretion, community autonomy, harm prevention, or participation.

Within these hybrid regimes, AI systems are often positioned not as replacements for human judgment, but as decision-support and workflow-shaping tools. They can triage content, automate high-confidence cases, and route ambiguous cases to human review. For example, Riehle et al. (2024) describe machine learning-based moderation as a configurable decision-support system in which automation depends on confidence thresholds, uncertain cases require manual moderation, and human moderators can overrule automated decisions. Such systems therefore do not merely detect harmfulness; they help structure which cases become visible to human moderators, which cases are presumed publishable, and which cases are treated as requiring restriction.

Professional moderation work, meanwhile, is shaped by institutional priorities and organizational constraints. Ruckenstein and Turunen (2020) show how commercial moderators often imagine automation as a way to offload repetitive "cleaning" tasks to machines so that humans can reorient toward higher-order "care" work, such as cultivating discussion culture. At the same time, professional moderation is embedded in policy compliance, organizational accountability, and risk management in ways that ordinary user judgments are not. This means that professional moderators should not be treated as a neutral gold standard, but as one governance actor whose decisions reflect a particular institutional role.

Beyond this technical and organizational division, the legitimacy of platform governance is often argued to depend not only on outcomes, such as removal or retention, but on how decisions are made and experienced as justified. Suzor et al. (2018) frame legitimacy as a normative "meta-concept" for evaluating the exercise of power by intermediaries against human-rights and procedural governance values. Katsaros et al. (2024) similarly argue that procedural justice concerns, such as voice, neutrality, respectful treatment, and explanation, are central to whether moderation is experienced as legitimate. Algorithmic systems can offer operational advantages such as proactive detection and queue sorting, but they may also raise tensions around individualized consideration, transparency, and reason-giving (Katsaros et al., 2024; Molina & Sundar, 2022).

A further complication is that many platforms also embed participatory elements into governance, from user reporting to broader forms of crowd input. Work on Reddit moderation, for instance, highlights crowdsourced flagging as a scalable mechanism for surfacing problematic content and notes that such signals can also provide moderators with perceived legitimacy to act, while also documenting limitations and vulnerability to misuse (Bozarth et al., 2023). Related empirical work on Reddit's adoption of bot moderators suggests that algorithmic tools can reconfigure human moderation practice rather than simply replacing it, altering how moderators police, explain, and suggest appropriate behavior in communities (He et al., 2025).

Once uncivil comments are understood as norm-relevant moderation categories, alignment becomes a governance question rather than a simple measure of correctness. The issue is not whether users, AI, or professional moderators are "right" in the abstract, but whether hybrid decision pipelines operationalize judgments closer to community thresholds, automated triage thresholds, professional adjudication, or some combination of these. This yields the article's second RQ:

> RQ2: How do users' publication decisions align with the platform AI system's publish/hold dispositions and with professional moderators' publish/block decisions?

2.3. Where Divergence Should Concentrate in Hybrid Moderation

Even when platforms adopt shared policy categories, moderation decisions routinely diverge across decision-makers because many cases require interpretive rather than mechanical judgment. A trade-off-centered view of moderation highlights that such divergence is structurally expected: Platforms balance competing values such as harm prevention and freedom of expression, and these balances are implemented through different actions, such as removal, labelling, downranking, or routing to review, as well as through broader governance philosophies (Jiang et al., 2023). Closely related work on moderation dilemmas emphasizes that differences in enforcement can reflect principled disagreement about which

errors are more costly, such as removing permissible speech versus failing to mitigate harm, rather than simple mistakes (Kozyreva et al., 2023).

Divergence may also concentrate in particular content domains because the five categories studied here differ in normative basis and likely enforcement threshold. Some categories, such as Personal Attacks, are closer to interpersonal disrespect or rudeness, where users may disagree about whether the comment crosses the line from harsh expression into a removable violation. Other categories, such as Hate Speech or Violence, may be perceived as more severe because they implicate safety, dignity, or intolerance. Doxxing raises privacy and exposure concerns, while Sexual Explicitness may depend strongly on platform context, audience expectations, and policy boundaries. Prior work suggests that such distinctions matter: intolerant comments are generally perceived as more offensive, more socially harmful, and more deletion-worthy than uncivil comments, whereas merely rude or harsh comments do not always produce the same reactions (Kümpel & Unkel, 2023; Rossini, 2022).

A second reason divergence may concentrate in particular items is context sensitivity. Intent is frequently policy-relevant yet difficult to infer reliably from short text, especially without sufficient contextual information (Wang et al., 2025), and providing conversational context can improve the quality and consistency of hate-speech annotation (Ljubešić et al., 2023). Work on humor as an online safety issue illustrates this clearly: The same utterance can function as benign joking, targeted harassment, or coded hate depending on context, and platform interventions often operate under limited contextual information (Matamoros-Fernández et al., 2023). In practice, this context dependence is also relevant for automated systems, where detection and routing decisions often rely on incomplete signals and must operate under constraints of scale and uncertainty (Arora et al., 2024; Riehle et al., 2024).

Divergence can also emerge from the organizational and procedural conditions under which moderation takes place. Professional moderation is embedded in institutional settings oriented toward policy compliance, consistency, and risk management, while user judgments are not made under the same accountability pressures (Ruckenstein & Turunen, 2020). At the same time, AI-enabled moderation is often deployed as workflow support, such as triage and routing, which can standardize treatment of borderline cases in ways that differ from professional adjudication or community judgment (He et al., 2025; Riehle et al., 2024). Procedural legitimacy perspectives further suggest that disagreements are consequential not only because they occur, but because they can be experienced as failures of neutrality, respectful treatment, or reason-giving—dimensions that shape acceptance and contestation of governance decisions (Katsaros et al., 2024; Suzor et al., 2018).

We thus examine where disagreements concentrate in hybrid moderation regimes by presenting the third, exploratory RQ:

> RQ3: Where (by uncivil category and/or specific items) do alignment and divergence between users, AI, and professional moderators appear to be strongest or weakest?

2.4. Who Aligns With Whom?

Even when a platform applies a single policy framework, moderation judgments need not be uniform across users. From a governance perspective, heterogeneity is not simply noise: Moderation decisions often involve contested evaluations of what counts as uncivil, demeaning, threatening, or unacceptable, and these evaluations can vary across individuals and social contexts. Work on hate-speech judgments, for example, emphasizes that responses to hateful expression frequently draw on moral evaluation and perceived harm, which can plausibly generate individual differences in thresholds for restriction (Ieracitano et al., 2024).

In hybrid moderation regimes, this heterogeneity matters because alignment with different decision-makers can reflect more than agreement on isolated items. It can signal which standards of judgment users tend to share: those embodied in professional enforcement, those embodied in AI-supported workflow decisions, or neither. Related work operationalizes user–moderator alignment as multi-dimensional, distinguishing awareness of rules and enforcement from support for rules and their application (Koshy et al., 2023). Although the present study does not measure all of these dimensions, it similarly treats alignment as a way to examine the relationship between community judgments and institutional moderation decisions.

Procedural legitimacy perspectives also suggest that user characteristics and platform experience may shape how people evaluate moderation decisions. Users who participate more actively in a community may have different expectations about acceptable speech than less engaged users, and demographic differences may influence perceptions of harm, offensiveness, or appropriate intervention. At the same time, because the present study compares decisions on a small set of real comments, we treat user-level predictors as exploratory rather than as the main theoretical test. These considerations motivate our fourth RQ:

> RQ4: To what extent do user characteristics, including demographics and platform engagement, predict alignment with AI and professional moderation decisions?

2.5. Justifications as Normative Reasoning

The study's final focus concerns how users justify their decisions. This emphasis on justification is consistent with governance perspectives that treat moderation as a site of value conflict and trade-offs, where decisions are rarely experienced as purely technical rule application. A trade-off-centered framework highlights that moderation choices routinely involve competing commitments, such as harm reduction versus participation and consistency versus contextual discretion, and that these value tensions shape enforcement outcomes (Jiang et al., 2023).

More concretely, moderation dilemmas often hinge on how people balance preventing harm against avoiding overreach or censorship. Accordingly, justifications can reveal whether users deny a violation, accept that a comment is problematic but treat it as below an enforcement threshold, or invoke higher-order constraints such as free speech (Kozyreva et al., 2023). This distinction is important because disagreement over moderation does not necessarily mean disagreement over whether a comment is objectionable. Users may agree that a comment is rude, offensive, or harmful, but disagree about whether removal is the appropriate response.

Legitimacy-oriented accounts further emphasize that reason-giving and perceived procedural fairness matter for how governance decisions are evaluated and contested. Moderation interventions are easier to accept when they can be connected to intelligible norms and processes, and harder to accept when they appear arbitrary or unaccountable (Katsaros et al., 2024; Suzor et al., 2018). In uncivil content domains, justifications are especially informative because judgments can be morally charged and context-sensitive, increasing disagreement about what norm is relevant and whether it has been violated.

These considerations motivate our final RQ:

RQ5: How do users justify their publish/block decisions when evaluating uncivil comments?

3. Methods

3.1. Study Design and Setting

We analyze a moderation dataset from AOL's comment sections, a US-based publisher/news-portal environment using OpenWeb's commenting infrastructure. OpenWeb operates a community and conversation platform for major publishers to support higher-quality discussion on publishers' own sites rather than external social media platforms. In this setting, comment sections are not treated simply as open posting spaces, but as publisher-adjacent conversational environments in which moderation is used to reduce low-quality, abusive, or norm-violating contributions and support more meaningful discussion.

The study combines (a) production moderation decisions generated within OpenWeb's moderation workflow by both AI and professional moderators and (b) ex-post user judgments about the same comments collected via an online survey. Each participating user evaluated the same set of real comments, indicated whether each comment should remain published or be removed, and provided explanations for the decision.

This study relies on a production-linked secondary dataset provided by OpenWeb. The survey instrument and data collection were designed and executed by the company, with input from the researcher during the planning stage, but not under full researcher control. The researcher received the full dataset after collection and conducted the analyses reported here independently. The study should therefore be understood as a secondary analysis of company-generated data rather than as a fully researcher-administered survey.

3.2. Materials and Comment Selection

The survey, executed by OpenWeb, included 16 real comments selected in consultation with the research team as typical examples intended to represent the theoretical constructs in the civil–epistemic norm-violation typology (Barnoy, Bakbani-Elkayam, et al., 2025; Keren et al., 2025). Two comments were selected for each of eight categories: five uncivil categories and three epistemic categories. The present study focuses on the five uncivil categories: Violence, Doxxing, Hate Speech, Sexual Explicitness, and Personal Attacks, yielding 10 comments, two per category. The six remaining comments were excluded from the present analysis because the study focuses on civil, interpersonal, and safety-related moderation categories rather than epistemic norm violations.

Comments were presented as text only, without usernames, thread context, timestamps, or article headlines. This design choice mirrors a common constraint in moderation and annotation research, where out-of-context judgments can differ systematically from in-context judgments (Ljubešić et al., 2023). The comment texts and associated metadata are available in the Supplementary File.

3.3. Moderation Decisions

For each focal comment, we observe two production-stage moderation decisions. First, the AI outcome captures the platform's first-stage workflow disposition. OpenWeb described the system as a proprietary, self-developed moderation tool that uses multiple signals and external API partners, including Google Perspective, as part of its moderation infrastructure. The present study does not evaluate the technical performance of a standalone AI model. Instead, it analyses the observed workflow dispositions produced by the platform's operational AI-assisted moderation system.

In the analytic dataset, the AI outcome is treated as a binary workflow orientation: publish-oriented versus approval-oriented/restrictive. Publish-oriented comments were those marked as "publish and moderate," meaning that the comment was published and routed for human verification by a professional moderator. Approval-oriented/restrictive comments were those requiring human approval before publication. The analytic dataset includes only comments for which both an AI workflow disposition and a professional moderation decision were available.

Second, professional moderation captures the final adjudication of each comment. Each comment received a single final decision by a professional paid moderator trained by OpenWeb, recorded as "approved" (publish) or "blocked" (remove/do not publish). Professional moderators were not part of the 863-user survey sample; their decisions were production-stage adjudications recorded separately from the ex-post user survey.

For analyses comparing AI, professionals, and users, we operationalize the AI moderation state as a binary tendency: AI-publish orientation versus AI-hold/restrictive orientation. This coding reflects whether the automated workflow initially treated the comment as publishable by default or as requiring approval/restriction before publication.

3.4. User Sample and Recruitment

Registered OpenWeb users were recruited via an email invitation. Participation was voluntary and uncompensated, and respondents could skip items and exit the survey at any point. A total of 892 responses were recorded at the survey entry stage. An initial "readiness" item asked whether respondents intended to proceed. Twenty-nine cases that did not proceed were excluded, leaving 863 unique users in the analytic sample.

Users were categorized by the platform into an ordinal engagement-level variable reflecting platform-derived behavioral status: Non-engaged, Reader, Engager, and Commenter. In this study, all four categories refer to registered users, as only registered users could be contacted for survey participation.

3.5. Survey Procedure

After entry, respondents evaluated the 16 comments in randomized order. For each comment, the main prompt asked: "What do you think we should do with the comment?" Respondents selected one of two options: (a) leave it as it is on the website, or (b) remove it from the website. In the analyses below, "publish" refers to the first option, and "block" or "remove" refers to the second option. Respondents were not shown any information about the AI triage state, the professional moderation decision, or the comment category.

After each publish/remove decision, respondents were asked to indicate a justification. Depending on their moderation choice, they were shown a structured list of predefined justification options, with an optional open-ended "other" field.

3.6. Measures

The first measure was user decision. The primary user outcome is a binary decision. Skipped items were recorded as missing decisions. Across the 10 uncivil items, there are 8,630 potential user–comment decisions (863 users × 10 comments), of which 42 are missing, yielding $N = 8{,}588$ valid decisions for the focal analyses.

The second measure is professional moderation final decision, which was recorded as approved/blocked for each comment.

The third measure, AI moderation, is recorded as a binary workflow orientation: publish-oriented versus approval-oriented/restrictive.

The fourth measure is user characteristics. Demographic variables (age, gender, education, ethnicity) were self-reported. Platform-derived engagement level is recorded as a four-category ordinal measure (Non-engaged → Reader → Engager → Commenter), ranging from registered users with no recorded engagement to users who posted a comment or reply. Non-engaged users are registered users with no recorded engagement with the platform. Readers are active readers of the conversation, defined as users who spent more than five seconds reading. Engagers are users who liked, disliked, or shared a comment. Commenters are users who posted a comment or reply. These categories therefore indicate increasing levels of participation in the platform environment.

The last measure is justifications. After making a publish or block decision, respondents were asked to provide a justification using predefined response options, with an optional open-ended "other" field. Accordingly, RQ5 analysis is based primarily on structured justification responses, supplemented where relevant by the smaller set of open-ended answers.

3.7. Analytic Strategy

We report descriptive statistics for user publish/block decisions overall and by uncivil category, along with item-level variation within categories (RQ1). For RQ2, we compare alignment between users and (a) AI moderation and (b) professional moderation. Alignment is assessed using pairwise agreement metrics and user-level paired comparisons. For each user, we compute their agreement rate with the AI and with

professional moderators across the 10 focal items. We then estimate the within-user difference using a paired bootstrap confidence interval and compare the number of users who aligned more with the AI versus more with professional moderators using a sign test, treating equal-alignment cases as ties. As a robustness check, we estimate a logistic model predicting user publish decisions from AI triage orientation and professional moderation decisions, with standard errors clustered by user.

For RQ3, we descriptively locate where alignment and divergence concentrate across the uncivil categories and specific items, including cases where AI and professional moderation disagree and cases where users diverge from both. For RQ4, we compute three user-level outcomes: (a) agreement rate with AI, (b) agreement rate with professional moderators, and (c) the difference between these agreement rates. We then model these outcomes as a function of user characteristics, including demographics and engagement. Finally, for RQ5, we analyze users' justification responses descriptively by decision type (publish/block). Because the justification item consisted primarily of predefined response options, the analysis focuses on the frequency distribution of these categories. Open-ended "other" responses are used secondarily and interpretively to illustrate how users elaborated or departed from the structured options.

4. Findings

4.1. User Moderation Distribution (RQ1)

Across the uncivil comment set, users chose to leave comments published in 5,678 of 8,588 decisions (66.12%), and to remove in 2,910 decisions (33.88%).

User decisions vary substantially across the five uncivil categories. The highest publication rate appears for Doxxing (users chose to publish 84.34% of the time; 1,443/1,711), followed by Personal Attacks (69.71%; 1,197/1,717) and Violent Behavior (66.47%; 1,146/1,724). Publication rates are lower for Sexually Explicit content (55.96%; 962/1,719) and Hate Speech (54.16%; 930/1,717).

Because each category is represented by only two comments, category-level patterns partly reflect item-specific differences. This dispersion is visible within categories: For example, Violent Behavior ranges from 82.39% publish for one item (711/863) to 50.52% publish for the other (435/861). Similar within-category spreads appear for Personal Attacks (83%/56.41%), Sexually Explicit (45.41%/66.55%), and Doxxing (90.91%/77.73%). Accordingly, the category-level differences above should be read as descriptive patterns in this comment set, rather than as stable category effects.

Across engagement types, publication rates are broadly similar: engagers publish 67.73%, non-engaged users 67.52%, commenters 64.35%, and readers 64.05%—suggesting modest baseline differences in leniency/strictness at this descriptive level.

To summarize RQ1, users published roughly two-thirds of uncivil comments in this set, with meaningful variation across categories and substantial within-category dispersion across the specific items.

4.2. Alignment of User, AI, and Professional Moderators' Decisions (RQ2)

Overall, users opted to publish 66.12% of comments, compared to 69.98% for the AI system's publish-oriented versus hold/restrictive disposition and 50.07% for professional moderators' publish versus block decision. Users' decisions aligned more closely with the AI system than with professional moderators. User–AI general agreement was 64.67% (Cohen's $\kappa = 0.189$), compared to user–professional agreement of 54.31% ($\kappa = 0.086$). Although both κ values indicate low absolute agreement, the relative pattern is consistent with the percentage-agreement results: Users' decisions were closer to the AI workflow disposition than to professional moderators' final adjudications.

The underlying confusion patterns reflect the same asymmetry. For users versus AI, users agreed with AI on 4,327 publish decisions and 1,227 hold/block decisions, with 3,034 disagreements. For users versus professional moderators, agreement consisted of 3,027 publish decisions and 1,637 block decisions, with 3,924 disagreements.

Because each user evaluated nearly all comments, we also treated users as the unit of inference and compared, within user, the proportion of decisions that matched the AI with the proportion that matched professional moderators. The mean user–AI agreement rate was 0.65, compared to 0.55 for user–professional agreement, yielding a mean within-user difference of 0.103. A paired user-level bootstrap confidence interval indicated that this difference was consistently positive, 95% CI [0.092, 0.114]. At the individual level, 523 users aligned more with the AI than with professional moderators, 222 aligned equally with both, and 118 aligned more with professional moderators. A sign test among non-tied users showed that AI-leaning alignment was more common than professional-leaning alignment ($p < 0.001$).

Finally, as a robustness check, we modelled users' publish decisions as a function of the AI publish orientation and the professional moderator publish decision, with standard errors clustered by user. Both predictors were positively associated with users' decisions, but the AI association was substantially stronger: AI publish orientation, OR = 2.22, 95% CI [2.07, 2.38], $p < 0.001$; professional moderator publish decision, OR = 1.25, 95% CI [1.17, 1.35], $p < 0.001$. A direct coefficient contrast confirmed the difference, $z = 10.94$, $p < 0.001$.

To summarize RQ2, users' publication decisions were systematically closer to the AI system's publish/hold orientation than to professional moderators' publish/block decisions.

4.3. Alignment Across Uncivility Categories (RQ3)

Because each uncivil category is represented by only two comments (10 items total), we treat RQ3 as descriptive/exploratory and focus on identifying where agreement and disagreement concentrate across categories and specific items.

Across the 10 uncivil comments, the AI system and professional moderators reached the same publish/block decision for six items (60%), and disagreed on four items (40%). Disagreements are not evenly distributed across categories: they occur in Doxxing (1/2 items), Sexually Explicit (1/2 items), and Hate Speech (2/2 items), while Personal Attacks (2/2 items) and Violence (2/2 items) show full AI–professional concordance.

In three of the four AI–professional disagreement items, users' majority decisions are closer to the AI decision than to the professional decision. For example, when the AI recommends publication but professionals block, users still tend to publish: 77.7% for a Doxxing item, 66.6% for a Sexually Explicit item, and 58% for a Hate Speech item. The remaining disagreement case is a Hate Speech item in which users are essentially split (50.3% publish), yielding near-equal alignment with the AI (block) and professionals (publish). These item-level patterns help locate the empirical source of the overall asymmetry documented in RQ2.

Even among the six items where AI and professionals agree, user decisions are not uniform. In one Sexually Explicit item that both AI and professionals published, users published only 45.4% of the time (i.e., users were more restrictive). Conversely, for a Personal Attack item and one Violence item that both AI and professionals blocked, users nevertheless published 56.4% and 50.5% of the time, respectively (i.e., users were more permissive or split). Thus, while AI–professional agreement is common, it does not guarantee convergence with community judgments.

To summarize RQ3: In most disagreement cases, users' majority decisions align more closely with the AI than with professional moderators, but users also diverge meaningfully from both actors on some items where AI and professionals agree.

4.4. User Characteristics Predicting Alignment (RQ4)

To examine user-level predictors of alignment, we modelled each respondent's relative AI alignment across the 10 uncivil comments. This measure was calculated as the user's agreement rate with the AI minus the user's agreement rate with professional moderators; positive values therefore indicate closer alignment with the AI, negative values indicate closer alignment with professional moderators, and zero indicates equal alignment with both.

We then tested whether demographics (self-reported age, gender, education, ethnicity) and engagement type predicted this relative alignment measure. In a parsimonious model including age and engagement type, age was associated with slightly weaker relative AI alignment ($\beta = -0.00115$ per year, 95% CI [−0.00204, −0.00026], $p = 0.011$), while engagement type differences were small and not statistically reliable. Consistent with this, age correlated negatively with user–AI agreement ($r = -0.11$) but was essentially unrelated to user–professional agreement ($r = 0.01$), indicating that the age pattern is primarily driven by weaker user–AI alignment among older users.

When adding the full set of demographic controls (gender, education, ethnicity) alongside engagement, the overall explanatory power remained very small ($R^2 = 0.01$, 0.03, depending on specification), and no demographic pattern emerged that was both robust and substantively large.

Accordingly, these results are best interpreted as bounded, secondary evidence: User characteristics explain only a small portion of variance in alignment, and the main empirical takeaway remains the overall alignment asymmetry documented in RQ2.

4.5. Users' Moderation Justifications (RQ5)

Users provided a justification for the vast majority of uncivil-item decisions (98%). These justification responses were primarily drawn from predefined survey options, supplemented by a smaller number of open-ended "other" responses. Below, we summarize the justification patterns separately for publish and block decisions.

Among publish decisions (N = 5,678), the most common justifications were norm-threshold oriented. A total of 38.2% selected or stated (in near-formulaic terms) that the comment was "not toxic" (e.g., "the comment is not toxic," "there is nothing wrong here," and so on), and 34.8% acknowledged incivility while rejecting removal as disproportionate (e.g., "toxic but not enough to justify removal"). A substantial minority invoked explicit principled toleration/anti-censorship considerations (7.8%), typically in the language of free speech (e.g., "free speech," "1st Amendment right," "freedom of speech").

Smaller groups referenced the comment as merely an opinion (2.1%), noted context uncertainty (1.5%; e.g., "no context/unknown context"), or proposed alternative remedies (0.6%; e.g., downvoting or ignoring rather than removal). These patterns suggest that users' publish decisions are often justified either by denying a civil norm violation or by treating the violation as falling below a removal threshold, with a non-trivial subset framing removal as an illegitimate form of censorship.

Among block decisions (N = 2,910), justifications most often took the form of a violation label. A total of 84.4% consisted of one of the five pre-defined uncivil-category terms. As a descriptive perception check, 51.2% of all block decisions and 60.7% of those using one of the five uncivil-category labels matched the assigned stimulus category, indicating moderate, uneven alignment between intended category and perceived reason for removal.

While a small category, it is worth mentioning that 6.4% of block justifications used epistemic labels ("untrue," "misleading," "conspirative") despite the stimulus set being primarily uncivil, indicating that some users interpret at least some uncivil items through an epistemic lens. The remaining 7.6% contained diverse evaluative language (e.g., "crude," "racist," "childish," "adds nothing to the conversation").

5. Conclusion

5.1. Different Moderation Actors, Different Governance Functions

This study shows that, in this real-world uncivil-comment dataset, users, AI-assisted triage, and professional moderators operationalized moderation thresholds differently. Across the 10 uncivil comments, users chose to publish roughly two-thirds of comments, the AI workflow disposition was similarly publication-oriented, and professional moderators were substantially more restrictive. This asymmetry also appeared at the user level: Most respondents aligned more closely with the AI system than with professional moderators. These findings suggest that in hybrid moderation systems, an important governance question is not only whether content is filtered effectively, but whose standards are being enacted when authority is distributed across users, automated systems, and professional reviewers (Jiang et al., 2023; Koshy et al., 2023; Suzor et al., 2018).

At the same time, greater user–AI alignment should not be read as evidence that AI moderation is more legitimate, more accurate, or normatively preferable. Similarity to user judgment is only one possible evaluative benchmark. Legitimacy-oriented work makes clear that platform governance must also be assessed in relation to procedural values such as transparency, accountability, neutrality, and the intelligibility of decision processes (Katsaros et al., 2024; Molina & Sundar, 2022; Suzor et al., 2018). Nor should professional moderation be treated as an objective gold standard. Professional moderators operate within institutional settings and policy mandates that differ from ordinary user judgment, including obligations to apply platform rules consistently and to manage reputational, legal, and safety-related risk.

This is especially important because the categories examined here differ in normative and practical stakes. Some content may be unlawful or clearly safety-relevant, while other content may be legally permissible but still inappropriate, uncivil, or unsuitable for a specific platform or publisher context (Galli et al., 2023). Professional moderators' more restrictive pattern may therefore reflect not simply greater sensitivity to incivility, but a different governance function: applying institutional thresholds under conditions of accountability. By contrast, user judgments may better indicate community tolerance thresholds, and AI-assisted triage may function as a scalable workflow mechanism that sorts content according to configured risk signals. The practical implication is therefore not that one actor is generally superior, but that different actors may be better suited to different moderation functions: AI for scalable triage, professional moderators for policy- and risk-sensitive adjudication, and user or community signals for understanding local expectations and tolerance.

5.2. Where Disagreement Concentrates

A second finding is that divergence was not evenly distributed across the uncivil categories. AI and professional moderators disagreed on four of the 10 items, with disagreement concentrated in both Hate Speech items, one Doxxing item, and one Sexually Explicit item. In three of these four disagreement cases, users' majority judgments were closer to the AI than to the professional moderator decision. At the same time, even when AI and professional moderators agreed, users did not always converge with them: In some items, users were more restrictive than both, while in others they were more permissive or split. The observed asymmetry is therefore not reducible to a simple "users versus professionals" divide. Rather, disagreement appears to concentrate in specific content domains and specific items.

This pattern is consistent with prior work suggesting that moderation disputes often intensify where context, ambiguity, and interpretive flexibility are especially salient. Judgments can change substantially when contextual information is missing, and the same utterance can be interpreted differently depending on conversational cues, perceived intent, humor, or target (Ljubešić et al., 2023; Matamoros-Fernández et al., 2023). In the present study, comments were presented without usernames, thread context, article context, or interaction history. Under such stripped-down conditions, borderline items may become especially vulnerable to divergent readings. This helps explain why disagreement concentrated in particular items rather than appearing uniformly across all categories.

5.3. User Reasoning and Heterogeneity

The justification findings add an important interpretive layer because they show not only how users voted, but also which kinds of reasons they selected or articulated. Because these justifications were captured primarily through structured response options, the findings should be interpreted as patterns in endorsed reasoning rather than as a fully open-ended map of users' spontaneous normative discourse. Still, the contrast between publish and block justifications is informative. Among publish decisions, users most often either denied that the comment violated a relevant norm at all ("not toxic") or acknowledged some problem while treating removal as disproportionate ("toxic but not enough"). A smaller but non-trivial subset framed removal as censorship or as a violation of free speech. Among block decisions, by contrast, users usually justified removal by labelling the comment as a violation category, such as Personal Attack, Hate Speech, Sexually Explicit, Violent Behavior, or Doxxing.

This asymmetry suggests that permissive and restrictive decisions may rely on different normative grammars. Publication is often defended through threshold reasoning and principled toleration, whereas blocking is often justified through straightforward recognition of a rule violation. This distinction matters because disagreement over moderation does not necessarily mean disagreement over whether a comment is objectionable. Users may recognize that a comment is rude, offensive, or problematic, but still disagree that it crosses the threshold for removal. Future moderation research should therefore distinguish recognition of norm violation from preferred sanctioning response rather than treating them as equivalent.

The user-level models further suggest that the overall alignment asymmetry was not strongly driven by easily identifiable demographic or engagement-based subgroups. Although older users showed slightly less alignment with the AI, the effect was small, and the explanatory power of the models remained weak. Engagement differences were also modest and statistically unreliable. This indicates that the user–AI alignment pattern was relatively broad within the sampled population, while also cautioning against strong claims about demographic predictors of moderation preferences.

5.4. Limitations and Future Research

Several limitations should qualify the interpretation of these findings. First, the uncivil analysis is based on only 10 comments, with just two items per category. The category-level patterns should therefore be treated as descriptive tendencies in this specific comment set rather than stable estimates of how users, AI systems, and professional moderators would judge these categories more generally. Future studies should examine larger and more systematically varied comment samples, including variation in severity, target, ambiguity, and contextual cues.

Second, the study relies on a production-linked secondary dataset generated by OpenWeb. Although the researcher had input into the study design and received the full dataset, the questionnaire implementation, recruitment, and data collection were managed by the company. This creates the usual constraints of secondary-data analysis: some measurement choices, response formats, and field procedures could not be fully standardized or revised post hoc. The justification measure, for example, was primarily structured rather than fully open-ended, limiting the extent to which the study captures spontaneous user reasoning.

Third, the comments were evaluated in a decontextualized format. Users saw text only, without thread or article context. This design makes the comparison across users, AI, and professional moderation more controlled, but it also differs from many real-world moderation situations. Future work should test whether the same alignment patterns hold when fuller contextual cues are available.

Fourth, the user sample consisted of registered OpenWeb users who voluntarily participated in the survey, so the findings should not be treated as a general estimate of public opinion about moderation. The results are best understood as evidence from a US-based, production-linked publisher-comment environment rather than from the general population, especially because norms around privacy exposure, doxxing, sexual explicitness, hate speech, and acceptable incivility may vary across contexts.

Finally, the AI variable captures a first-stage workflow disposition within OpenWeb's moderation system, not the final decision of a fully automated moderation process. The system is proprietary and combines OpenWeb's own moderation infrastructure with external signals, including API partners such as Google Perspective. The study therefore cannot isolate the performance of a specific AI model. Moreover, because the analysis includes comments for which both AI workflow dispositions and professional moderation decisions were available, it cannot estimate agreement in cases filtered out before human review or cases that would never enter a comparable human adjudication pathway.

Overall, the findings suggest that the key issue in AI-assisted moderation is not simply whether uncivil content is detected, but whose standards are enacted when moderation authority is shared across community judgment, automated triage, and professional review. In this dataset, users were consistently closer to the AI's publish versus hold orientation than to professional moderators' final decisions. That does not settle how moderation ought to be organized. It does, however, clarify that hybrid moderation systems should be evaluated not only by whether they make decisions at scale, but by which actor's judgment they implement, under what conditions, and for which moderation purpose.

Acknowledgments

The author wishes to acknowledge the invaluable contributions of the late Eran Ayalon. His collaboration during his time at OpenWeb was instrumental to the early stages of this research, and this article is dedicated to his memory.

Funding

Publication of this article in open access was made possible through the institutional membership agreement between Erasmus University Rotterdam and Cogitatio Press.

Conflict of Interests

OpenWeb covered modest research expenses. The collaboration also informed OpenWeb's internal moderation work, but OpenWeb had no control over the academic work.

LLMs Disclosure

LLM (ChatGPT 5.2/4) was used for brainstorming, verification of analysis, and proofreading.

Supplementary Material

Supplementary material for this article is available online in the format provided by the author (unedited).

References

Arora, A., Nakov, P., Hardalov, M., Sarwar, S. M., Nayak, V., & Dinkov, Y. (2024). Detecting harmful content on online platforms: What platforms need vs. where research efforts go. *ACM Computing Surveys*, *56*(3), Article 72. https://doi.org/10.1145/3603399

Barnoy, A., Bakbani-Elkayam, S., Miller, B., & Keren, A. (2025). The role of epistemic norms in mitigating the spread of misinformation. *New Media & Society*. Advance online publication. https://doi.org/10.1177/14614448251385917

Barnoy, A., Freiman, O., Keren, A., & Miller, B. (2025). Norm violations in online discourse: Epistemic and civil foundations for platform design and moderation. *Social Epistemology*. Advance online publication. https://doi.org/10.1080/02691728.2025.2566124

Bormann, M., Tranow, U., Vowe, G., & Ziegele, M. (2022). Incivility as a violation of communication norms—A typology based on normative expectations toward political communication. *Communication Theory*, *32*(3), 332–362. https://doi.org/10.1093/ct/qtab018

Bozarth, L., Im, J., Quarles, C., & Budak, C. (2023). Wisdom of two crowds: Misinformation moderation on Reddit and how to improve this process—A case study of Covid-19. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW1), Article 155. https://doi.org/10.1145/3579631

Galli, F., Loreggia, A., & Sartor, G. (2023). The regulation of content moderation. In D. Moura Vicente, S. de Vasconcelos Casimiro, & C. Chen (Eds.), *The legal challenges of the fourth industrial revolution* (pp. 63–87). Springer. https://doi.org/10.1007/978-3-031-40516-7_5

He, Q., Hong, Y., & Raghu, T. S. (2025). Platform governance with algorithm-based content moderation: An empirical study on Reddit. *Information Systems Research*, *36*(2), 1078–1095. https://doi.org/10.1287/isre.2021.0036

Ieracitano, F., Balenzano, C., Girardi, S., Gemmano, C. G., & Comunello, F. (2024). Online hate speech as a moral issue: Exploring moral reasoning of young Italian users on social network sites. *Social Science Computer Review*, *42*(1), 25–47. https://doi.org/10.1177/08944393231161124

Jiang, J. A., Nie, P., Brubaker, J. R., & Fiesler, C. (2023). A trade-off-centered framework of content moderation. *ACM Transactions on Computer-Human Interaction*, *30*(1), Article 3. https://doi.org/10.1145/3534929

Katsaros, M., Kim, J., & Tyler, T. R. (2024). Online content moderation: Does justice need a human face? *International Journal of Human–Computer Interaction*, *40*(1), 66–77. https://doi.org/10.1080/10447318.2023.2210879

Keren, A., Barnoy, A., Freiman, O., & Miller, B. (2025). New Atlantis 2.0: Designing epistemically healthy online conversation. In P. J. Connolly, S. C. Goldberg, & J. M. Saul (Eds.), *Conversations online: Explorations in philosophy of language* (pp. 337–356). Oxford Academics. https://doi.org/10.1093/oso/9780198872085.003.0016

Koshy, V., Bajpai, T., Chandrasekharan, E., Sundaram, H., & Karahalios, K. (2023). Measuring user-moderator alignment on r/ChangeMyView. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW2), Article 286. https://doi.org/10.1145/3610077

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*(7), Article e2210666120. https://doi.org/10.1073/pnas.2210666120

Kümpel, A. S., & Unkel, J. (2023). Differential perceptions of and reactions to incivil and intolerant user comments. *Journal of Computer-Mediated Communication*, *28*(4), Article zmad018. https://doi.org/10.1093/jcmc/zmad018

Ljubešić, N., Možetič, I., & Kralj Novak, P. (2023). Quantifying the impact of context on the quality of manual hate speech annotation. *Natural Language Engineering*, *29*, 1481–1494. https://doi.org/10.1017/S1351324922000353

Masullo Chen, G., Muddiman, A., Wilner, T., Pariser, E., & Stroud, N. J. (2019). We should not get rid of incivility online. *Social Media + Society*, *5*(3). https://doi.org/10.1177/2056305119862641

Matamoros-Fernández, A., Bartolo, L., & Troynar, L. (2023). Humour as an online safety issue: Exploring solutions to help platforms better address this form of expression. *Internet Policy Review*, *12*(1). https://doi.org/10.14763/2023.1.1677

Molina, M. D., & 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), Article zmac010. https://doi.org/10.1093/jcmc/zmac010

Riehle, D. M., Wolters, A., & Müller, K. (2024). Adopting artificial intelligence in a decision support system: Learnings from comment moderation systems. *Procedia Computer Science*, *239*, 1847–1855. https://doi.org/10.1016/j.procs.2024.06.366

Rossini, P. (2020). Beyond toxicity in the online public sphere: understanding incivility in online political talk. In W. H. Dutton (Ed.), *A research agenda for digital politics* (pp. 160–170). Edward Elgar. https://doi.org/10.4337/9781789903096.00026

Rossini, P. (2022). Beyond incivility: Understanding patterns of uncivil and intolerant discourse in online political talk. *Communication Research*, *49*(3), 399-425.. https://doi.org/10.1177/0093650220921314

Ruckenstein, M., & Turunen, L. L. M. (2020). Re-humanizing the platform: Content moderators and the logic of care. *New Media & Society*, *22*(6), 1026–1042. https://doi.org/10.1177/1461444819875990

Suzor, N., Van Geelen, T., & Myers West, S. (2018). Evaluating the legitimacy of platform governance: A review of research and a shared research agenda. *International Communication Gazette*, *80*(4), 385–400. https://doi.org/10.1177/1748048518757142

Wang, X., Koneru, S., Venkit, P. N., Frischmann, B., & Rajtmajer, S. (2025). The unappreciated role of intent in algorithmic moderation of abusive content on social media. *Harvard Kennedy School Misinformation Review*, *6*(3). https://doi.org/10.37016/mr-2020-180

Yu, Z., Otto, L., Assenmacher, D., & Wagner, C. (2024). A systematic review of the effects of AI-assisted moderation on individuals and groups. *Human-Machine Communication*, *9*, 167–188. https://doi.org/10.30658/hmc.9.10

About the Author

Aviv Barnoy is an assistant professor of media and communication at Erasmus University Rotterdam. His research bridges academic theory and professional practice in trust, crisis communication, strategic communication, and platform governance, with particular attention to how institutions communicate and govern under contested conditions.