societies

MDPI

Article

Negotiating Free Speech and Content Moderation: User Reactions to Meta's Community Notes on Facebook and Instagram

Antonia Cava [1], Andrea Nucita [1], Assunta Penna [1] and Debora Maria Pizzimenti [2,*]

[1] Department of Cognitive Sciences, Psychology, Education, and Cultural Studies, University of Messina, 98122 Messina, Italy; antonia.cava@unime.it (A.C.); andrea.nucita@unime.it (A.N.); assunta.penna@unime.it (A.P.)
[2] Department of Humanities, University of Calabria, 87036 Arcavacata di Rende, Italy
* Correspondence: debora.pizzimenti@unical.it

Abstract

Social media platforms are increasingly reshaping content moderation through participatory governance models. In January 2025, Meta announced that its third-party fact-checking program in the United States would be replaced by Community Notes, raising questions concerning platform governance, freedom of expression, epistemic authority, and misinformation. This study examines how commenters interpreted this policy change through a comparative qualitative content analysis of comments posted in response to Mark Zuckerberg's official Facebook and Instagram announcements. Following research authorization, we examined 1000 top-level comments per post through the Meta Content Library research environment. After applying the eligibility criteria, the analytical corpora comprised 1535 English-language comments: 895 from Facebook and 640 from Instagram. Eligible comments were manually coded according to evaluative stance, argument type, and communication tone. Opposition to Community Notes was more frequent in the Facebook corpus (61.7%) than in the Instagram corpus (43.4%). Facebook comments more frequently provided explicit ideological, epistemic, technological, and comparative justifications, whereas Instagram contained a substantially larger proportion of comments without an explicit argument (43.6%, compared with 11.5% on Facebook). Constructive communication was the predominant tone in both discussion environments (60.0% on Facebook and 58.8% on Instagram), and the proportions of the three communication tones were similar in the two coded corpora. Overall, the study provides a situated account of how commenters interpreted and negotiated collaborative moderation and platform governance across two specific discussion environments.

Keywords: Community Notes; content moderation; platform governance; misinformation; qualitative content analysis

Check for updates

Academic Editors: Felice Addeo and Gabriella Punziano

Received: 8 July 2026
Revised: 28 September 2026
Accepted: 30 September 2026
Published: 4 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Rather than functioning merely as communication technologies, digital platforms have evolved into complex socio-technical infrastructures through which individuals access information, participate in public debate, construct identities, and negotiate shared meanings [1–3]. Therefore, scholarship on digital societies has increasingly emphasized that platforms not only mediate social interactions but also actively participate in organizing social life, structuring public discourse, and producing new forms of knowledge

and power [4,5]. Within this broader transformation, the concept of platformization has emerged to describe the progressive expansion of platform infrastructures, economic processes, and governance frameworks across multiple sectors of society [6]. Research on platform agency also examines how users' resources and digital literacies shape participation within algorithmically structured environments [7]. As platforms have become central infrastructures of communication, decisions regarding the visibility, circulation, and legitimacy of online information have progressively shifted from traditional public institutions toward privately governed digital ecosystems [8,9]. Platform governance has consequently become a defining characteristic of contemporary digital societies, with content moderation constituting one of its central mechanisms. Beyond the removal of illegal or harmful content, moderation increasingly regulates participation, visibility, and the boundaries of acceptable public discourse [5,8]. This governance involves human moderation work [10] as well as automated systems that raise technical and political challenges [11]. Growing concerns regarding online misinformation, political polarization, and harmful content (particularly following major electoral campaigns, the COVID-19 pandemic, and international geopolitical crises) have further intensified expectations that platforms should actively intervene to protect the quality of online information while simultaneously preserving freedom of expression [12,13]. The concept of information disorder provides a framework for distinguishing different forms of problematic information rather than treating all misleading content as a single category [14]. Content moderation has therefore become a key arena in which competing values, including safety, transparency, democratic participation, and free speech, are continuously negotiated. Recent regulatory developments, including the European Union's Digital Services Act, have further reinforced expectations concerning platform transparency, accountability, and systemic-risk governance [15]. At the same time, major technology companies have begun reconsidering the role of centralized moderation itself, gradually experimenting with more participatory governance models involving users directly in content evaluation. A particularly significant example of this transformation occurred in January 2025, when Meta announced the gradual replacement of its third-party fact-checking program in the United States with *Community Notes*, a collaborative moderation system modeled on the system previously implemented on X (formerly Twitter). The initiative aims to promote greater freedom of expression while reducing perceived bias associated with professional fact-checking [16]. Unlike traditional fact-checking, the Community Notes model relies on user-generated contextual annotations intended to provide additional information about online content. However, although moderation responsibilities appear partially redistributed to platform communities, the visibility of annotations remains governed by proprietary algorithms that evaluate contributor reliability and cross-perspective consensus. From a governance perspective, Community Notes has been presented as a model that combines distributed participation with algorithmic decision-making [16]. This development provides an opportunity to examine how users interpret evolving approaches to content moderation and the broader governance principles underpinning them. Accordingly, this study adopts a sociological perspective by focusing on how commenters interpret this transition and the meanings they attribute to issues such as expertise, trust, freedom of expression, and the role of platforms in regulating online information. Whereas professional fact-checking relied primarily on institutional expertise, the transition to Community Notes has been presented as increasing users' participation in the evaluation of online information. At the same time, these participatory practices remain embedded within algorithmically governed infrastructures. The existing research has examined the effectiveness of Community Notes in limiting misinformation and their effects on trust in fact-checking [17–19]. While these studies provide important evidence regarding the performance of collaborative moderation, less attention has been devoted to

how commenters interpret this governance model, negotiate its legitimacy, and connect it to broader tensions concerning freedom of expression, misinformation, institutional trust, and algorithmic governance. Building on these considerations, this article investigates how Facebook and Instagram commenters interpret Meta's transition from third-party fact-checking to Community Notes across two distinct platform environments. More specifically, it examines commenters' evaluative positions, the argumentative types through which those positions are justified, and the communicative tones through which they are expressed. Drawing on comments examined and manually coded through the Meta research environment, the study adopts a comparative qualitative content analysis. Its contribution is primarily empirical and interpretative, shifting attention from the technical performance of Community Notes to the situated meanings through which commenters publicly evaluate and contest this governance intervention.

Accordingly, the study addresses the following research questions:

RQ1. How do Facebook and Instagram commenters evaluate Meta's transition from third-party fact-checking to Community Notes?

RQ2. What types of arguments do commenters mobilize in discussing this transition, and how do these types differ across the two observed platform environments?

RQ3. What communicative tones characterize commenters' responses to Meta's policy change, and how do they differ across the two observed platform environments? The remainder of this article is organized as follows. The next section reviews the literature on content moderation, collaborative governance, and Community Notes. The subsequent section presents the research design, describing the data collection procedures and analytical methods adopted in this study. The Section 4 reports the qualitative findings concerning commenters' evaluative stances, argument types, and communication tones across the two discussion environments. The Discussion interprets these findings in relation to contemporary debates on platform governance, collaborative moderation, and epistemic authority. Finally, the article summarizes the principal contributions of the study, discusses its limitations, and outlines directions for future research.

2. Theoretical Framework

The rapid growth of digital communication has exposed several limitations of third-party fact-checking as one specific mechanism within broader systems of content moderation. In this study, content moderation is used as the broader governance category, while third-party fact-checking refers specifically to the institutional verification of potentially misleading content by external fact-checking organizations. Community Notes is treated as a participatory annotation system situated within this broader landscape of platform moderation and information governance. Although third-party fact-checking has constituted an important mechanism within contemporary content-moderation systems, recent research has highlighted several structural challenges that reduce their effectiveness in large-scale digital environments. First, professional fact-checking faces important issues of scalability. The speed of online communication creates challenges for timely fact-checking and community-based responses to misleading content [20]. Questions of public trust also arise in evaluations of fact-checking: experimental research has examined how community notes and other corrective formats affect trust in fact-checks [17]. These limitations have stimulated growing interest in participatory and community-driven forms of moderation and information evaluation, in which platform users actively contribute to contextualizing online content. Collaborative moderation is not entirely new within digital environments. X (formerly Twitter) introduced Birdwatch in 2021, subsequently renamed Community Notes, as a community-based approach to adding contextual annotations to potentially misleading content [20]. Rather than relying exclusively on third-party fact-checking or-

ganizations, the system enables eligible users to write contextual annotations that are subsequently evaluated by other contributors. Importantly, however, publication does not depend on simple majority voting but on proprietary algorithms designed to identify agreement among contributors with different viewpoints, thereby combining collective participation with algorithmic governance [16]. Recent research has examined both the potential and the limitations of Community Notes as a form of community-based information evaluation. Experimental evidence suggests that contextual annotations may be perceived as more trustworthy than simple misinformation flags, although these effects depend on how the intervention is presented and should not be generalized beyond the populations and conditions examined [17]. Research has also shown that community-based and professional fact-checking processes may select different claims for verification, indicating that participatory annotation does not merely reproduce professional fact-checking at a larger scale but can generate different patterns of attention and coverage [21].

Evidence concerning effectiveness is similarly mixed. Some studies have found that the display of Community Notes can reduce engagement with and diffusion of false or misleading information [18], whereas other analyses indicate that delays in producing and displaying notes may limit their capacity to intervene during the earliest and most viral stages of content circulation [19]. The perceived helpfulness of notes is also influenced by their evidentiary characteristics, including whether they cite sources that are regarded as politically unbiased [22]. At the same time, patterns of target selection and political asymmetry remain contested: recent research indicates that political groups may be flagged at different rates, while emphasizing that these differences can reflect variation in misinformation-sharing behavior rather than bias in the annotation system itself [23].

These findings show that Community Notes cannot be evaluated solely in terms of scale or aggregate participation. Its operation depends on which claims contributors select, how rapidly notes are produced, which sources are considered credible, and which annotations satisfy the platform's cross-perspective consensus requirements. From an analytical perspective, Community Notes should therefore not be understood as functionally equivalent to third-party fact-checking, even where it is introduced as its institutional replacement. It represents a governance arrangement involving community participation, platform rules, and algorithmic selection [16,24]. Beyond questions of effectiveness, scalability, and information quality, Community Notes raises a broader issue concerning how epistemic authority is distributed and perceived in platform environments. Unlike third-party fact-checking, in which judgments are institutionally associated with professional verification organizations, Community Notes distributes the production and evaluation of contextual annotations among eligible contributors. However, this redistribution does not transfer authority entirely from institutions to users. Platform-defined eligibility criteria, rating procedures, consensus requirements, and ranking algorithms continue to determine which contributions become publicly visible.

Recent peer-reviewed research supports understanding Community Notes as a hybrid socio-technical governance arrangement rather than as a straightforward democratization of fact-checking. Matamoros-Fernández and Jude conceptualize Community Notes as an infrastructure of "soft moderation" and show that its data structures and design priorities affect which forms of harm can be recognized and addressed [24]. Research on moderation legitimacy similarly demonstrates that perceptions of platform governance are shaped by public claims concerning bias, fairness, institutional responsibility, and the accessibility of mechanisms through which moderation decisions may be challenged [25]. Evidence that contextual community annotations can increase trust in fact-checking further suggests that legitimacy depends not only on who produces a corrective intervention but also on its form, explanatory content, and perceived evidentiary basis [17].

Against this peer-reviewed foundation, Augenstein et al.'s recent preprint can be used more narrowly to describe Community Notes as reflecting an emerging epistemology of collaborative fact-checking, in which credibility judgments are produced through interactions among contributors, algorithms, and platform institutions [26]. The preprint therefore extends, rather than independently supports, the argument developed here. Comparatively less attention has been devoted to how commenters themselves interpret these changing arrangements of authority and connect them to questions of expertise, trust, misinformation, freedom of expression, and platform legitimacy. Examining such interpretations contributes to understanding the situated social meanings attached to Community Notes beyond assessments of its technical performance. Because commenters' interpretations are embedded within specific platform environments, platform culture and affordances provide relevant contextual concepts for the comparison of Facebook and Instagram. Digital platforms can be understood as networked publics whose technical features and socially sedimented conventions organize, without deterministically producing, possibilities for interaction and public expression [27–29]. Instagram has been specifically examined as a visual social media environment in which images, videos, hashtags, memes, and emojis form part of its communicative practices [30]. These characteristics provide contextual grounds for comparing the two discussion environments but do not establish in advance that their commenters will display different evaluative, argumentative, or communicative patterns.

The two environments may also differ in audience composition and patterns of use. Recent population-level data indicate differences in the age distribution of Facebook and Instagram users in the United States [31]. However, these aggregate data do not establish the demographic characteristics of the commenters included in the present corpus. Platform affordances, audience composition, post format, engagement volume, comment visibility, and retrieval order are therefore treated as possible contextual factors rather than as demonstrated causal explanations for the differences observed in the analysis. The study adopts platform governance, situated within the broader process of platformization, as its primary analytical lens. Recent research further indicates that the legitimacy of content moderation is not determined exclusively by the formal characteristics of governance mechanisms. It is also constructed and contested through public interpretations of bias, fairness, institutional responsibility, and the opportunities available to users and creators to challenge platform decisions [25]. This perspective is particularly relevant to the present study because Meta's transition to Community Notes was publicly framed as a response to perceived bias in third-party fact-checking. The analysis therefore examines whether and how commenters reproduced, challenged, or reformulated this legitimacy claim within the two observed discussion environments.

Within this framework, epistemic authority and platform culture are treated as complementary concepts that help specify, respectively, how authority over information is negotiated and how evaluative and communicative practices are expressed within the two observed discussion environments. The analysis therefore examines how a specific governance intervention becomes an object of public evaluation, how commenters justify their positions through different argumentative types, and how these evaluations are expressed through different communicative tones. The contribution of the study lies in providing a situated empirical account of the evaluative stances, argument types, and communication tones through which commenters interpreted and contested Meta's transition to Community Notes across two specific discussion environments, rather than in proposing a general theory of platform governance.

3. Materials and Methods

Building upon the theoretical framework presented above, this study adopts a comparative qualitative content analysis [32,33] to investigate comments responding to Meta's transition from third-party fact-checking to Community Notes. The analysis examines evaluative stance, argument type, and communication tone in two specific discussion environments. Table 1 summarizes the characteristics of the two announcements, their approximate post-level engagement metrics at the time of collection, and the number of comments included in the initial working corpus.

Table 1. Characteristics of the two Meta announcements analyzed in this study. Approximate post-level engagement figures were read from the Meta Content Library web interface and recorded on 2 June 2026; they were not calculated from the sampled comments. These figures describe engagement displayed at the time of collection, not engagement when the comments were posted. On the same date, 1000 displayed top-level comments per post were selected for manual eligibility assessment and coding.

Platform	Post Format	Likes/ Reactions	Comments (Total)	Shares/Reposts	Saves	Comments in the Initial Working Corpus
Facebook	Video post	1.1M	599K	167K	26.3K	1000
Instagram	Reel	470K	75K	233K	N/A [1]	1000

Note. K = thousand; M = million. [1] N/A = not available. Instagram save counts were not publicly displayed in the Meta Content Library web interface when the metrics were recorded.

The two posts were selected because they constituted Mark Zuckerberg's official announcement of the policy change on the two Meta platforms and generated substantial user engagement. At the time the engagement metrics were recorded, the Facebook post had received approximately 599,000 comments and the Instagram Reel approximately 75,000 comments. This high level of participation made the announcement a significant case for examining how commenters publicly responded to and interpreted the transition to Community Notes. Facebook and Instagram were selected to compare responses to the same policy announcement across two platform environments within the same corporate ecosystem. Scholarship on platform affordances and visual social media provides a contextual basis for this comparison [27–30], without establishing in advance how commenters in the two corpora would articulate their responses. According to the Pew Research Center (2025) [31], 80% of U.S. adults aged 30–49 and 74% of those aged 50–64 reported using Facebook, whereas Instagram use was reported by 80% of adults aged 18–29 and 19% of those aged 65 and older. These population-level patterns provide contextual background for the comparison but do not establish the demographic characteristics of the commenters included in this study. Comparing commenters' responses across these environments therefore makes it possible to examine how the same governance intervention is interpreted and negotiated within different socio-technical contexts.

3.1. Qualitative Data Collection and Sampling

Systematic collection of the working corpora took place on 2 June 2026, following approval of access to the Meta Content Library and API on 26 May 2026. We used the Meta Content Library web interface to locate Zuckerberg's two announcement posts and view their comments. We set a target of 1000 top-level comments displayed for each post. Top-level comments were selected because the unit of analysis was a direct response to the announcement; replies and subsequent exchanges between commenters were outside the study's scope. The selected comments were inspected and manually assessed for

eligibility and, where eligible, coded within the authorized research environment. For each selected comment, the coder recorded an internal case label and an eligibility decision. For eligible comments, the coder also recorded the assigned categories for stance, argument type, and communication tone, together with brief descriptive notes on features relevant to those decisions. The interface offered date-range and chronological sorting options, but the settings governing the display order during selection were not documented. We therefore cannot characterize the selected comments as a chronological or randomized sample. An API-based chronological or randomized draw was not used to constitute these qualitative working corpora. For the final analysis, comments that did not meet the analytical eligibility criteria were excluded. In the Facebook working corpus, 105 of the initial 1000 comments were excluded: 53 were written in languages other than English and 52 were unrelated to Community Notes or the policy announcement, including generic greetings, statements of geographical origin, and other comments without a relevant connection to the topic under investigation. Among the 1000 Facebook comments in the initial working corpus, none consisted solely of emojis or other non-textual elements. Accordingly, no Facebook comment was excluded from the analytical corpus under this criterion. The resulting Facebook analytical corpus comprised 895 eligible comments. In the Instagram working corpus, 360 of the initial 1000 comments were excluded: 226 were classified as emoji-only or otherwise non-textual, 114 were written in languages other than English, and 20 were unrelated to the policy announcement or lacked sufficient interpretable content for coding. The resulting Instagram analytical corpus comprised 640 eligible comments. Excluded cases were retained in the working dataset but were not included in the final analytical corpus or its reported frequencies. The same eligibility criteria were applied to both working corpora. Brevity alone was not grounds for exclusion. Short comments were retained when their meaning could be interpreted in relation to the policy announcement, even if they did not articulate a justification; these were coded as "no explicit argument" when appropriate. Comments consisting solely of emojis or other elements that did not provide sufficient interpretable content for the three coding dimensions were excluded. An emoji accompanying an otherwise eligible comment did not, by itself, make that comment ineligible. For example, "😢😢" was excluded because it consisted solely of emojis, whereas "Well done 👏👏👏" was retained because its text expressed an interpretable evaluation of the policy announcement. The latter was coded as supportive instance, as containing no explicit argument, and as constructive in tone. For example, "4th!" was excluded as unrelated to the policy announcement, whereas "Excellent news!" was retained because it expressed an evaluation of the announcement without providing an explicit justification; the latter was coded as "no explicit argument". Sampling was not continued after these exclusions to replenish either analytical corpus to 1000 eligible comments. We reassessed the eligibility of comments in both working corpora using the same clarified criteria across all three coding dimensions. This review distinguished brief but interpretable responses, retained even when they contained no explicit argument, from non-textual or unrelated comments excluded from the final analysis. We also corrected documented errors in the analytical records and individual coding assignments. The resulting final analytical corpora comprise 895 Facebook and 640 Instagram comments, each coded in all three dimensions. Codes previously assigned to comments that remained eligible were generally retained, except where an individual correction was documented. A co-author reviewed borderline eligibility cases without access to the primary coder's earlier category assignments; the primary coder made the final eligibility decisions. This review was limited to borderline cases and did not constitute a blind reassessment of the complete corpus.

The non-probability sampling strategy was used to construct numerically balanced working corpora across the two discussion environments rather than to reproduce the proportional distribution of their full commenting populations. The post-level comment counts indicate different volumes of engagement, but they do not establish the size or composition of the retrievable top-level comment streams from which the working corpora were assembled. The comparison is consequently framed as a comparative case analysis of two specific discussion environments surrounding the same policy announcement. Absolute frequencies and within-corpus percentages describe the manually coded comments in each discussion environment and support the qualitative comparison. They are not estimates of the distributions in the complete comment streams or among Facebook and Instagram users more broadly. To address the research questions, the qualitative analysis adopted a coding strategy combining deductive and inductive procedures, consistent with established approaches to qualitative content analysis [32–34]. An initial coding framework was developed from the literature on platform governance, collaborative moderation, misinformation, epistemic authority, and freedom of expression [5,11,13]. Successive readings of the corpus enabled the analytical categories to be progressively refined as recurrent interpretative patterns emerged from users' comments through continuous comparison between empirical observations and theoretical concepts, following an iterative process of category development [32–34]. Manual coding facilitated continuous engagement with the empirical material throughout the analytical process, allowing the iterative refinement of categories while preserving the contextual interpretation of users' comments [32–34]. Each comment was treated as the unit of analysis and interpreted in relation to its textual content and the policy announcement [32,33]. The final analytical framework consisted of three complementary dimensions designed to capture different aspects of users' responses to Meta's policy change. The framework was designed to distinguish between evaluative orientation, argumentative justification, and communicative performance, thereby capturing complementary dimensions of users' responses to platform governance. The three dimensions were conceived as analytically distinct but complementary. Together, they enabled the reconstruction not only of commenters' evaluations of Community Notes but also of the reasoning through which these evaluations were justified and the communicative styles through which they were publicly expressed. To enhance methodological transparency and reproducibility, a detailed qualitative codebook was developed for the three analytical dimensions. For each category, the codebook specifies its operational definition, inclusion and exclusion criteria, distinctions from conceptually adjacent categories, and indicative lexical and semantic markers used to support contextual interpretation during manual coding. These markers were treated as interpretative cues rather than deterministic classification rules, as category assignment was based on the contextual meaning of each comment.

The first dimension, stance, identified the overall evaluative position expressed toward the introduction of Community Notes. Comments were classified as supportive, opposed, or neutral according to their overall orientation toward the policy change.

The second dimension, argument type, examined the principal form of reasoning through which commenters justified their evaluations. Six categories were identified: ideological, epistemic, experiential, technological, comparative, and no explicit argument. The latter category included comments expressing a clear evaluative position without providing an explicit justification or argumentative elaboration (e.g., brief expressions of approval, disapproval, or emotional reactions). The "no explicit argument" category therefore referred specifically to the absence of an articulated justification and was analytically distinct from communication tone. A comment could consequently be coded as containing no explicit argument while simultaneously displaying a constructive, adversarial, or ironic communicative tone. Thus, "no explicit argument" captured whether a justification was

present, whereas "constructive" captured how the comment was rhetorically expressed; the two codes referred to different analytical properties of the same comment.

Finally, communication tone captured the rhetorical style through which commenters expressed their reactions, independently of their evaluative stance. Comments were coded as constructive, adversarial, or ironic, allowing the analysis to distinguish between what position was expressed, how that position was justified, and the communicative style through which it was articulated. Stance, argument type, and communication tone were therefore coded as analytically distinct but complementary dimensions of the same comment [32,33]. Table 2 summarizes the analytical framework adopted for the qualitative content analysis, including the three coding dimensions, their categories, and their analytical purposes.

Table 2. Summary of the analytical framework adopted for the qualitative content analysis.

Analytical Dimension	Categories	Analytical Purpose
Stance	Supportive; Opposed; Neutral	Identifies the overall evaluative position toward Meta's introduction of Community Notes.
Argument type	Ideological; Epistemic; Experiential; Technological; Comparative; No explicit argument	Examines the principal form of reasoning through which commenters justify their evaluations.
Communication tone	Constructive; Adversarial; Ironic	Captures the rhetorical style through which commenters express their evaluations, independently of their stance.

The three dimensions were coded separately. Each eligible comment received one valid code for stance, one for argument type, and one for communication tone. "No explicit argument" was treated as a substantive argument-type category rather than as a missing value. The final analytical corpus comprised 895 Facebook comments and 640 Instagram comments, for a total of 1535 comments coded in all three dimensions. Absolute category frequencies and percentages were calculated separately for each platform, using 895 and 640 as the respective denominators. The distributions are presented descriptively and interpreted alongside the qualitative material; no inferential statistical tests are reported or used to support the findings in the present version.

3.2. Coding Procedure

The qualitative content analysis was conducted by a single primary coder. To support analytical consistency, transparency, and credibility, several methodological procedures were implemented [33]. First, a structured coding framework was developed for the three analytical dimensions—stance, argument type, and communication tone—with operational definitions and decision criteria specified for each category. The coding framework was refined through an initial pilot coding round conducted on a randomly selected subset of 100 eligible comments, comprising 50 comments from Facebook and 50 from Instagram. The pilot round was used to clarify category definitions, examine ambiguous cases, and refine the coding rules before the complete analytical corpus was coded. Second, an audit trail was maintained throughout the coding process to document coding decisions, ambiguous cases, and the rationale for category assignments, thereby supporting procedural transparency and consistency over time. Third, borderline and ambiguous cases were periodically discussed with the co-author as a form of peer debriefing, while the primary coder retained responsibility for the final coding decisions. The corpus was not independently double-coded; consequently, no intercoder reliability coefficient was calculated

or claimed. The reported frequencies and percentages therefore summarize the primary coder's classifications. They are used to describe patterns within the two analytical corpora, while the interpretation draws on the coded comments and illustrative qualitative material.

A detailed qualitative codebook was developed to guide the manual coding process. It specifies the analytical dimensions, operational definitions, inclusion and exclusion criteria, distinctions between adjacent categories, decision rules, and indicative lexical and semantic-pragmatic cues. It also contains constructed examples illustrating the application of the coding framework; these examples are not verbatim comments from the analytical corpus. The lexical and semantic-pragmatic indicators were treated as interpretative cues rather than deterministic classification rules, as category assignments were based on the contextual meaning of each comment.

The codebook and an aggregated frequency matrix are provided as Supplementary Materials. The codebook documents the coding dimensions, category definitions, eligibility criteria, and decision rules. The frequency matrix reports category counts by platform and contains no comment-level records. The audit trail and the underlying case-level analytical records are not included in the Supplementary Materials.

3.3. Ethical Considerations

The study analyzed exclusively publicly available comments posted in response to Mark Zuckerberg's official Facebook and Instagram announcements. The research team's application for access to the Meta Content Library and API was approved by an independent reviewer of the Center for the Advancement of Social Data (CASD) on 26 May 2026. Systematic data collection was conducted on 2 June 2026 at 10:00 CEST (UTC+2), following the procedure described in Section 3.1. Access to and use of the Meta Content Library and API were governed by the Meta Research Tools Terms and Conditions. No interaction with users took place during data collection. The researchers made no attempt to identify individual users, collect additional personal information, or access private content.

Consistent with the ethical recommendations of the Association of Internet Researchers (AoIR), publicly available comments were treated as context-dependent research material rather than as freely reusable data, and particular attention was devoted to minimizing potential risks to individual users [35]. All quotations reproduced in the Section 4 were de-identified by removing usernames and other directly identifying information. Only a limited number of excerpts necessary to illustrate the analytical categories were reproduced.

The study adhered to the principles of proportionality, contextual integrity, and responsible internet research throughout data collection, analysis, and reporting [35]. Data access and processing were conducted in accordance with the conditions governing Meta's research data-access infrastructure. A formal determination was requested from the institutional ethics committee. In a written determination dated 20 September 2026, the committee confirmed that prior ethical approval was not required because the study analyzed publicly available comments and involved no recruitment of, contact with, or interaction with users.

4. Results

The qualitative analysis first examined commenters' overall evaluative orientation toward Meta's introduction of Community Notes.

Table 3 reports the absolute frequencies for each coding category in the two analytical corpora.

Across both discussion environments, opposition represented the most frequent stance, although its distribution differed between the Facebook and Instagram cases examined (Figure 1).

Table 3. Observed frequencies of coded comments by platform, analytical dimension, and category. Each eligible comment received one code per dimension; category counts therefore sum to 895 for Facebook and 640 for Instagram within each dimension.

Coding Dimension	Category	Facebook (*n*)	Instagram (*n*)
Stance	Opposed	552	278
	Supportive	231	164
	Neutral	112	198
	Total: stance	895	640
Argument type	Ideological	338	168
	Epistemic	111	17
	Experiential	142	135
	Technological	125	21
	Comparative	76	20
	No explicit argument	103	279
	Total: argument type	895	640
Communication tone	Constructive	537	376
	Adversarial	199	138
	Ironic	159	126
	Total: communication tone	895	640

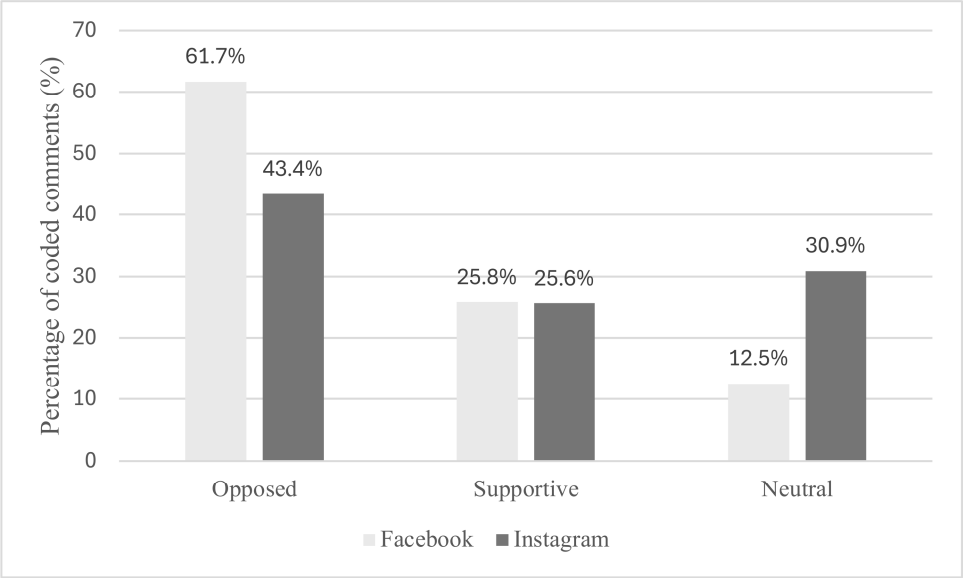


Figure 1. Distribution of commenter stance on Facebook (n = 895) and Instagram (n = 640). Percentages were calculated separately within each platform using the corresponding analytical corpus as the denominator.

Among the Facebook comments included in the analysis (n = 895), 61.7% were opposed to the policy change, 25.8% were supportive, and 12.5% were neutral. The Facebook discussion was therefore characterized by a clear predominance of critical reactions. Many commenters interpreted the replacement of third-party fact-checking as a politically motivated decision that would weaken mechanisms designed to counter misinformation. Several comments questioned Meta's credibility, describing the policy as a concession to political pressure or as a decision likely to increase the circulation of false information.

The Instagram corpus displayed a more heterogeneous distribution of evaluative positions. Among the Instagram comments included in the analysis (n = 640), 43.4% opposed Community Notes, 30.9% adopted a neutral position, and 25.6% expressed support for the new moderation model. Although opposition remained the largest category, supportive and neutral positions accounted for larger proportions of the Instagram corpus than of the Facebook corpus, resulting in a more balanced distribution across the three stance categories. These differences describe the two selected discussions and do not isolate an effect of platform. Supportive comments generally welcomed the reform as a step toward greater freedom of expression, whereas neutral comments frequently expressed uncertainty regarding the practical implementation of Community Notes without clearly supporting or opposing the initiative. This contrast is also visible in the qualitative material. A brief supportive reaction was "This is great news!" (Facebook; stance: supportive; argument type: no explicit argument; communication tone: constructive). Opposition was expressed in "Ah, lies will be allowed now—you have no shame" (Facebook; stance: opposed; argument type: epistemic; communication tone: adversarial). A comment called for procedural safeguards: "There should also be due process and appeal to counter censorship" (Facebook; stance: neutral; argument type: ideological; communication tone: constructive). It did not clearly endorse or reject the transition to Community Notes. Overall, opposition was the most frequent stance in both coded corpora. Its proportion was higher in the Facebook corpus, whereas neutral comments accounted for a larger share of the Instagram corpus. Supportive comments represented similar proportions in the two discussions.

Beyond commenters' overall evaluations, the qualitative analysis explored the argumentative patterns through which these positions were justified. Clear differences emerged between the two discussion environments (Figure 2).

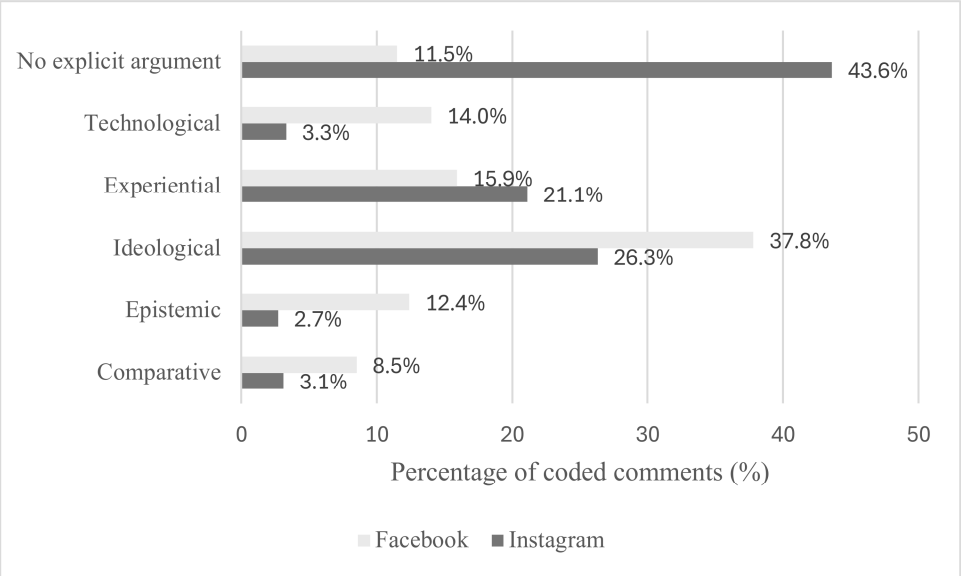


Figure 2. Distribution of argument types on Facebook (N = 895) and Instagram (N = 640). Percentages were calculated separately for each platform using its analytical corpus as the denominator.

In the Facebook corpus (n = 895), ideological arguments represented the largest category (37.8%), followed by experiential arguments (15.9%), technological arguments (14.0%), epistemic arguments (12.4%), comments without an explicit argument (11.5%), and comparative arguments (8.5%). Within the analyzed English-language Facebook corpus, the higher proportion of ideological arguments indicates that Community Notes was frequently interpreted through broader political and normative debates concerning freedom of expression, censorship, democracy, institutional trust, and political bias. Discussions

frequently extended beyond the technical characteristics of Community Notes to address wider questions concerning Meta's role in governing public discourse. Experiential arguments typically referred to commenters' previous experiences with content moderation, account restrictions, or perceived censorship. Technological and epistemic arguments were also relatively common, with commenters discussing algorithms, platform infrastructures, misinformation, truthfulness, and the legitimacy of professional fact-checking. Comparative arguments frequently evaluated Community Notes by contrasting the new system with Meta's previous moderation policies or with the Community Notes model implemented on X. The analyzed English-language Instagram corpus presented a different argumentative profile (n = 640). The descriptive distributions of argument types differed across the two coded corpora. Comments without an explicit argument represented a larger proportion of the eligible Instagram corpus, while explicit epistemic, technological, and comparative arguments represented larger proportions of the eligible Facebook corpus. Ideological arguments remained the largest explicit argument category on Instagram (26.3%), followed by experiential arguments (21.1%). The most distinctive finding was the large proportion of Instagram comments classified as containing no explicit argument (43.6%). These comments consisted primarily of brief evaluative statements, affective reactions, or expressions of approval or disapproval that did not provide an identifiable justification for the commenter's position. Comparative (3.1%), technological (3.3%), and epistemic arguments (2.7%) were comparatively rare. The argument-type categories are illustrated by comments from the analytical corpora. An ideological justification appears in "This is an effective step towards removing barriers and strengthening freedom of speech" (Facebook; stance: supportive; argument type: ideological; communication tone: constructive). An experiential justification linked the commenter's previous experience of moderation to explicit support for the announced change: "Ok Mark, as much as it pains me to give you a complete pass for suppressing conservative viewpoints, including me as a legitimate and thoughtful voice, I am glad to hear about a new direction for Meta. I will support the effort." (Facebook; stance: supportive; argument type: experiential; communication tone: constructive). An Instagram commenter explicitly disagreed with the announcement by referring to a previous experience of account restriction: "I don't agree. Once you put me in a Facebook prison for a picture taken in the museum so I couldn't call my dog!" (Instagram; stance: opposed; argument type: experiential; communication tone: constructive). A technological justification focuses on the operation of automated moderation: "Freedom it's not real in Meta. Your AI bans every phrase without understanding the meaning in context" (Facebook; stance: opposed; argument type: technological; communication tone: constructive). An epistemic justification appears in "I don't get it. Fact checking doesn't limit free speech. Fact checking just points out disinformation" (Facebook; stance: opposed; argument type: epistemic; communication tone: constructive). A comparative argument invoked the experience of Community Notes on another platform: "Yeah because community notes are going so well for Twitter. The sensible people have left in droves to go to Blue Sky to avoid the right wing nut jobs" (Facebook; stance: opposed; argument type: comparative; communication tone: ironic). By contrast, "Makes IG much better" expresses approval without providing an explicit justification (Instagram; stance: supportive; argument type: no explicit argument; communication tone: constructive).

To separate the presence of an explicit justification from the type of justification offered, we also examined the distributions after removing "no explicit argument" from both corpora. Among comments containing an explicit argument (792 Facebook comments and 361 Instagram comments), ideological arguments accounted for 42.7% and 46.5%, respectively; experiential arguments for 17.9% and 37.4%; epistemic arguments for 14.0% and 4.7%; technological arguments for 15.8% and 5.8%; and comparative arguments for 9.6%

and 5.5%. These percentages use only comments with an explicit argument as their denominators, whereas Figure 2 uses the full eligible corpora of 895 Facebook and 640 Instagram comments. Within this restricted descriptive comparison, experiential arguments occupied a larger share of the Instagram comments, while epistemic and technological arguments occupied larger shares of the Facebook comments. Taken together, these findings show different argumentative profiles across the two discussion environments. Facebook comments more frequently contained an explicit justification, whereas the Instagram corpus contained a markedly larger proportion of comments without an explicit argument. Relative to the full eligible corpora, ideological, epistemic, technological, and comparative arguments accounted for larger shares on Facebook, while experiential arguments accounted for a larger share on Instagram. Among comments with an explicit argument, however, ideological arguments represented similar proportions in the two cases. The third analytical dimension examined the communicative tone through which commenters expressed their evaluations (Figure 3).

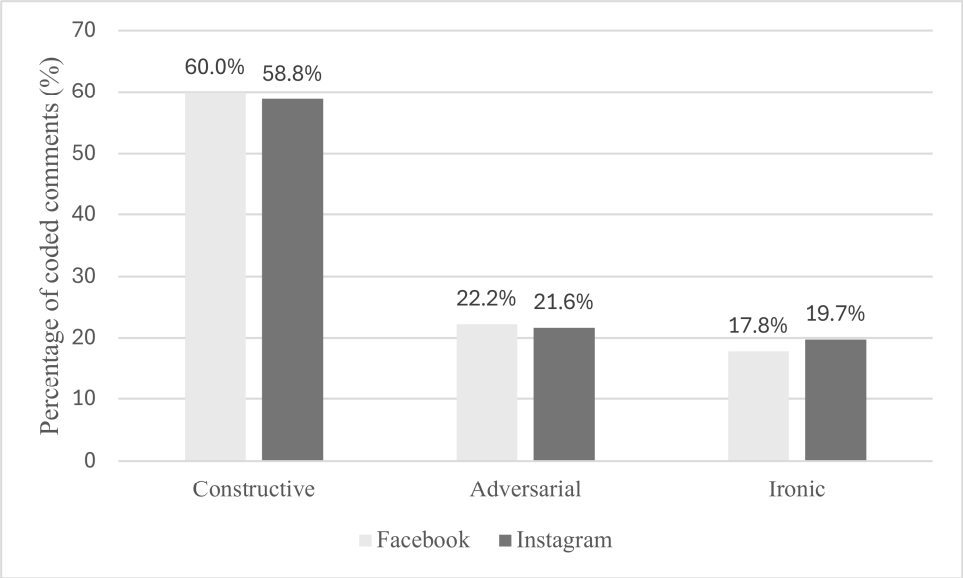


Figure 3. Communication tone across Facebook (n = 895) and Instagram (n = 640). Percentages were calculated separately within each platform using the corresponding analytical corpus as the denominator.

Constructive communication represented the most frequent tone in both discussion environments. Among the Facebook comments included in the analysis (n = 895), 60.0% were coded as constructive, 22.2% as adversarial, and 17.8% as ironic. Constructive comments expressed evaluations or disagreements without direct attacks or ridicule and could occur across different evaluative stances. Adversarial comments relied more heavily on direct accusations or attacks directed toward Meta, Mark Zuckerberg, political actors, or other relevant subjects, whereas ironic comments frequently expressed criticism through sarcasm, parody, or ridicule. A similar distribution emerged in the Instagram corpus (n = 640). Constructive comments accounted for 58.8% of the corpus, followed by adversarial comments (21.6%) and ironic comments (19.7%). The proportions of constructive, adversarial, and ironic comments were similar in the two coded corpora. Although the Instagram corpus contained a larger proportion of comments without an explicit argument, constructive communication remained the most frequent tone. The illustrative comments show how communication tone, stance, and argument type were coded separately. Constructive criticism included "Very disappointing. Especially to equate fact checking to censorship." (Facebook; stance: opposed; argument type: epistemic; communication tone: constructive).

A more adversarial criticism appeared in "What an absolutely shit choice. I hope you pay for it" (Facebook; stance: opposed; argument type: no explicit argument; communication tone: adversarial). Ironic criticism attributed the decision to political pressure: "Translation: I'm scared of Trump's threats to jail me for fact-checking his lies, so I'm going to let him lie all he wants. It's good for business in Trumpistan." (Facebook; stance: opposed; argument type: ideological; communication tone: ironic). Taken together, these findings indicate highly similar distributions of communication tone across the two cases. Constructive communication was the most frequent rhetorical form in both corpora, while adversarial and ironic communication accounted for smaller proportions of comments. In descriptive terms, the differences observed in stance and argument type were more pronounced than those observed in communication tone.

5. Discussion

This study provides a situated account of how English-language commenters in two specific discussion environments interpreted Meta's introduction of Community Notes. Rather than assessing the effectiveness of Community Notes as a moderation system, the analysis examined how commenters evaluated and discussed this policy change. Within the analyzed corpora, Community Notes emerged as a contested socio-technical governance arrangement through which questions of platform authority, participation, expertise, institutional trust, and freedom of expression were publicly negotiated [1,3,5,6]. These patterns suggest that Meta's announcement was not interpreted merely as a technical modification to content moderation. Many commenters situated the initiative within broader normative debates concerning censorship, political neutrality, democratic accountability, institutional credibility, and freedom of expression. Supportive commenters frequently interpreted the replacement of third-party fact-checking as an opportunity to reduce perceived ideological bias and expand freedom of expression. Opposed commenters, in contrast, often described the same policy as an abdication of platform responsibility or as a weakening of safeguards against misinformation. Neutral comments, which were particularly prevalent on Instagram, more often expressed uncertainty or requested information about how Community Notes would operate in practice.

The findings are consistent with scholarship conceptualizing content moderation as a central component of platform governance rather than as a merely technical process of removing, ranking, or labeling content [5,11]. Within the two discussions, Community Notes functioned as a symbolic object through which commenters articulated competing expectations concerning the role of platforms in governing online public discourse. Governance innovations consequently acquire social meaning not only through their technical design but also through the political, cultural, and experiential frameworks within which commenters interpret them [4,6,36]. The central difference in argumentative patterns concerns whether commenters articulated an explicit justification for their position. Restricting the descriptive comparison to comments that did provide a justification reveals a second pattern: experiential arguments represented a larger share of the Instagram comments with an explicit argument, whereas epistemic and technological arguments represented larger shares of the Facebook subset. The two observations should therefore be interpreted separately, particularly because eligibility decisions disproportionately removed non-textual material from the Instagram working corpus. Ideological arguments constituted the largest argumentative category on Facebook and the largest explicit argumentative category on Instagram. These comments connected Community Notes to wider debates about censorship, democracy, freedom of expression, political bias, and institutional trust. The prominence of such arguments indicates that commenters frequently evaluated the initia-

tive through broader political and normative frameworks rather than exclusively in terms of its anticipated technical performance.

Experiential arguments were also important in both corpora, particularly on Instagram. These comments typically drew on personal encounters with content moderation, account restrictions, content removal, or perceived censorship. Such accounts show how evaluations of new governance arrangements may be shaped by commenters' previous experiences with platform regulation. Community Notes was therefore interpreted not in isolation but against a background of accumulated perceptions concerning Meta's past moderation practices.

Although epistemic arguments represented a smaller proportion of the corpus than ideological or experiential arguments, they addressed a theoretically important question: how authority to determine the reliability of information should be distributed among professional fact-checkers, platform users, algorithms, and Meta itself. Commenters did not uniformly interpret Community Notes as a democratization of credibility. Some regarded community participation as a means of reducing centralized or expert control, whereas others understood it as a potentially unreliable transfer of responsibility from professional fact-checkers to platform users.

These findings complement research on the effectiveness and perceived credibility of collaborative fact-checking systems [17–19]. The present analysis draws attention instead to how commenters interpret the redistribution of epistemic authority associated with the transition from third-party fact-checking to community-based moderation. Within the discussions examined, the policy change was interpreted as a hybrid arrangement involving users, algorithms, professional expertise, and institutional platform power. Community participation may decentralize some evaluative functions, while Meta continues to exercise substantial authority through the design of contributor eligibility, ranking procedures, aggregation mechanisms, and note-visibility rules.

The large proportion of Instagram comments without an explicit argument should not be interpreted as evidence of an absence of evaluative engagement. Many of these comments communicated approval, disapproval, surprise, or other affective reactions without providing an explicit justification. The finding instead indicates that the two analyzed discussions differed in the extent to which evaluations were discursively elaborated. Facebook comments more frequently articulated an explicit basis for their position, whereas Instagram contained more concise and affective responses. One possible explanation for the different argumentative profiles is the composition of the audiences reached by the two announcements. Although both posts communicated the same policy change and came from the same public figure, one was a Facebook video post and the other an Instagram Reel; their followings and patterns of exposure may have differed. Political views, prior experiences of moderation, and demographic characteristics of the commenters were not measured, so their contribution to the observed distributions cannot be separated from post format, comment visibility, or interface display order. Our interpretation is therefore confined to what commenters articulated in these two corpora: it identifies different forms of justification in the observed discussions without attributing those forms to an inherent property of either platform.

The proportions of constructive, adversarial, and ironic comments were similar in the two coded corpora. Thus, although the corpora differed in stance and argumentative profile, they displayed highly similar distributions of communication tone.

This result confirms the analytical importance of distinguishing stance, argument type, and communication tone. Opposition to Community Notes was not necessarily communicated through hostility, personal attacks, or ridicule. Commenters could criticize Meta's decision while maintaining a constructive tone, just as supportive or neutral positions

could be expressed ironically or adversarial. Similarly, the absence of an explicit argument did not necessarily imply an adversarial or disruptive communicative style. Evaluative orientation, argumentative justification, and rhetorical form therefore captured distinct aspects of public responses to the policy announcement.

Taken together, the findings highlight an underlying tension within collaborative moderation. Community Notes was presented as a means of increasing participation and addressing concerns surrounding centralized fact-checking. However, the analyzed reactions indicate that participation does not automatically produce perceived legitimacy or trust. Commenters continued to raise concerns about political bias, misinformation, institutional credibility, algorithmic neutrality, and platform power. Rather than resolving controversies surrounding content moderation, the transition to Community Notes appeared, within these cases, to reorganize those controversies around a different governance model.

At the same time, the coexistence of supportive, opposed, and neutral positions cautions against interpreting Community Notes exclusively as a source of distrust. For some commenters, community participation represented a legitimate form of decentralization and a potential corrective to perceived bias in professional fact-checking. For others, it represented the withdrawal of institutional responsibility or the substitution of professional expertise with an uncertain collective process. These interpretations are not mutually reducible and demonstrate that the legitimacy of collaborative moderation is socially negotiated rather than inherent in the technological mechanism itself.

From a methodological perspective, the study shows how qualitative content analysis can distinguish evaluative stance, argumentative justification, and communication tone within the same set of comments. Counts and percentages describe the primary coder's classifications within the two eligible corpora and support their qualitative interpretation. Because the comments were not independently double-coded, intercoder reliability was not assessed. The descriptive distributions are not representative estimates of the wider comment streams.

6. Conclusions

Across the two analyzed discussions, commenters expressed different evaluative positions and drew on different forms of justification when interpreting Meta's transition to Community Notes. The findings show the value of examining stance, argumentative justification, and communication tone separately. In the discussions examined, the shift toward a participatory moderation model did not, by itself, secure commenters' trust in or acceptance of Meta's governance decision. In these discussions, commenters assessed Meta's authority and the proposed role of community contributors in light of their differing expectations and prior experiences of moderation.

The findings also demonstrate the importance of examining moderation systems not only in terms of their technical design or measurable effectiveness but also as objects of public interpretation and contestation. Within the cases examined, commenters connected Community Notes to broader questions concerning credibility, freedom of expression, political bias, expertise, institutional responsibility, and platform authority. The study therefore provides a situated account of how collaborative moderation was publicly interpreted in two specific discussions rather than evidence of a general transformation in platform legitimacy or epistemic governance.

The analysis also has broader societal implications. The examined discussions illustrate how content moderation, algorithmic governance, and epistemic authority become matters of public negotiation. Strengthening citizens' understanding of the socio-technical mechanisms through which platforms organize visibility, credibility, and participation consequently represents an important dimension of contemporary media education and digital

citizenship [37,38]. Such understanding may support more informed and reflective participation in digital public spaces across formal, non-formal, and informal learning contexts.

Several limitations must be considered. First, the study examined a non-probability corpus of English-language comments drawn from one announcement on each platform. Following the application of the eligibility criteria, the analyzed corpus comprised 895 Facebook comments and 640 Instagram comments. The exclusion rates differed across the two working corpora. Exclusions of emoji-only or otherwise non-textual comments were concentrated in the Instagram working corpus. The two analytical corpora therefore represent not only different proportions of their initial samples but also differently constituted subsets of those samples. The findings consequently describe the eligible English-language comments retained under the study's coding criteria and should not be generalized to the complete discussions or to the broader populations of Facebook and Instagram users.

Second, comments were selected in the order displayed by the Meta research interface at the time of collection. We did not verify how that order was determined, and we did not select comments through a chronological or randomized draw. Sampling was not continued to replace excluded comments or to obtain 1000 eligible observations per platform. The language composition of the complete comment streams was not established; therefore, the extent to which the interface display order affected the language composition of the working corpora cannot be estimated from these data. The observed differences may also reflect audience composition, post format, engagement volume, visibility mechanisms, and other contextual factors that the comparative case design cannot disentangle. We did not record the commenters' account identifiers in the working dataset. Consequently, we cannot determine from these data whether the same account contributed more than one top-level comment. The reported frequencies count comments, not distinct accounts.

Third, the analysis focused on a single announcement on each platform. Although the posts communicated the same policy change and were published by the same public figure, the study cannot isolate platform as the causal explanation for the observed differences. The findings must therefore be interpreted as case-specific, platform-associated patterns rather than as general effects of Facebook or Instagram.

Fourth, the qualitative content analysis was conducted by a single primary coder. The pilot coding exercise, structured codebook, audit trail, and peer debriefing procedures supported consistency and transparency but do not constitute a formal assessment of intercoder reliability. Some degree of interpretative subjectivity consequently remains, particularly in the classification of ambiguous or borderline comments.

Finally, automated sentiment analysis was not retained in the final research design. Lexicon-based and automated models may have difficulty identifying sarcasm, irony, negation, and context-dependent meanings, particularly without independent validation against human-coded data. The study therefore relied on manual qualitative coding. Future research could combine validated computational methods with qualitative analysis to examine larger and more diverse corpora.

Further research should extend the comparative design to additional policy announcements, time periods, languages, and platform environments. Longitudinal analysis could assess whether public interpretations change after users gain direct experience with Community Notes. Interviews with platform users, Community Notes contributors, professional fact-checkers, moderators, and platform designers could also provide a deeper understanding of how collaborative moderation is interpreted and experienced from different positions within contemporary platform ecosystems.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/soc16100324/s1.

Author Contributions: Conceptualization, A.C., A.N., A.P. and D.M.P.; Methodology, A.N., A.P. and D.M.P.; Software, A.N.; Formal analysis, D.M.P.; Investigation, A.C., A.N., A.P. and D.M.P.; Data curation, A.N. and D.M.P.; Writing—original draft, A.P. and D.M.P.; Writing—review & editing, A.C. and D.M.P.; Supervision, A.C. and A.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This manuscript has obtained an ethical exemption based on the following considerations: ① Regarding the research content: The data analyzed in this study were derived from publicly available comments posted on Facebook and Instagram. The study does not involve human subjects research in the traditional sense, nor does it involve any intervention or interaction with individuals. ② Regarding the nature of the research: This study adopts an observational and non-interventional research design, focusing on the analysis of publicly accessible social media comments from Facebook and Instagram. No experimental procedures, interventions, or direct engagement with users were involved. ③ Regarding the research methodology: The study did not involve participant recruitment, direct contact, or interaction with users, nor did it access any private user communications. Data were accessed through the Meta Content Library web interface in accordance with Meta's authorized research procedures. The analysis reported in this manuscript consisted of manual qualitative coding of publicly available comments. All potentially identifiable information, including usernames, was removed from the qualitative dataset, and all quotations presented in the manuscript were anonymized to protect user privacy. ④ Regarding institutional confirmation: Furthermore, the Chair of the Ethics Committee at COSPECS, the authors' affiliated institution, reviewed the relevant documentation concerning this study and determined that the research did not require prior ethical approval. Therefore, formal ethical review was not deemed necessary.

Informed Consent Statement: Informed consent was not required because the study analyzed publicly available comments in accordance with the ethical recommendations of the Association of Internet Researchers (AoIR) and Meta's research policies.

Data Availability Statement: The codebook and aggregated frequency matrix are available in the Supplementary Materials. The underlying case-level analytical records are not publicly available because access to and use of the source comments are subject to Meta's research-data conditions and applicable privacy requirements. The Supplementary Materials contain no case-level records.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Couldry, N.; Hepp, A. *The Mediated Construction of Reality*; John Wiley & Sons: Cambridge, UK, 2018.
2. van Dijck, J.; Poell, T.; De Waal, M. *The Platform Society: Public Values in a Connective World*; Oxford University Press: Oxford, UK, 2018.
3. Couldry, N.; Mejias, U.A. *The Costs of Connection: How Data Are Colonizing Human Life and Appropriating It for Capitalism*; Stanford University Press: Stanford, CA, USA, 2019.
4. Beer, D. The social power of algorithms. *Inf. Commun. Soc.* **2017**, *20*, 1–13. [CrossRef]
5. Gillespie, T. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*; Yale University Press: New Haven, CT, USA, 2018.
6. Poell, T.; Nieborg, D.; van Dijck, J. Platformisation. *Internet Policy Rev.* **2019**, *8*, 1–13. [CrossRef]
7. Punziano, G.; Trezza, D. Agency in social media platforms: The role of capitals and literacies under algorithmic environments. *Inf. Commun. Soc.* 2026, *ahead of print*. [CrossRef]
8. Gorwa, R. What is platform governance? *Inf. Commun. Soc.* **2019**, *22*, 854–871. [CrossRef]
9. Balkin, J.M. Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation. *UC Davis Law Rev.* **2018**, *51*, 1149–1210. [CrossRef]
10. Roberts, S.T. *Behind the Screen: Content Moderation in the Shadows of Social Media*; Yale University Press: New Haven, CT, USA, 2019. [CrossRef]
11. Gorwa, R.; Binns, R.; Katzenbach, C. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data Soc.* **2020**, *7*, 1–15. [CrossRef]

12. Vosoughi, S.; Roy, D.; Aral, S. The spread of true and false news online. *Science* **2018**, *359*, 1146–1151. [CrossRef] [PubMed]
13. Suzor, N.P. *Lawless: The Secret Rules That Govern Our Digital Lives*; Cambridge University Press: Cambridge, UK, 2019.
14. Wardle, C.; Derakhshan, H. *Information Disorder: Toward an Interdisciplinary Framework for Research and Policymaking*; Council of Europe: Strasbourg, France, 2017. Available online: https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html (accessed on 28 June 2026).
15. European Parliament; Council of the European Union. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and Amending Directive 2000/31/EC (Digital Services Act). *Off. J. Eur. Union* **2022**, *L277*, 1–102. Available online: https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng (accessed on 28 June 2026).
16. Meta. More Speech and Fewer Mistakes. 2025. Available online: https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/ (accessed on 28 June 2026).
17. Drolsbach, C.P.; Solovev, K.; Pröllochs, N. Community notes increase trust in fact-checking on social media. *PNAS Nexus* **2024**, *3*, gae217. [CrossRef] [PubMed]
18. Slaughter, I.; Peytavin, A.; Ugander, J.; Saveski, M. Community notes reduce engagement with and diffusion of false information online. *Proc. Natl. Acad. Sci. USA* **2025**, *122*, e2503413122. [CrossRef] [PubMed]
19. Chuai, Y.; Tian, H.; Pröllochs, N.; Lenzini, G. Did the Roll-Out of Community Notes Reduce Engagement With Misinformation on X/Twitter? *Proc. ACM Hum.-Comput. Interact.* **2024**, *8*, 428. [CrossRef]
20. Wirtschafter, V.; Majumder, S. Future challenges for online, crowdsourced content moderation: Evidence from Twitter's community notes. *J. Online Trust Saf.* **2023**, *2*, 1–11. [CrossRef]
21. Pilarski, M.; Solovev, K.O.; Pröllochs, N. Community notes vs. snoping: How the crowd selects fact-checking targets on social media. *Proc. Int. AAAI Conf. Web Soc. Media* **2024**, *18*, 1262–1275. [CrossRef]
22. Solovev, K.; Pröllochs, N. References to unbiased sources increase the helpfulness of community fact-checks. *Sci. Rep.* **2025**, *15*, 25749. [CrossRef] [PubMed]
23. Renault, T.; Mosleh, M.; Rand, D.G. Republicans are flagged more often than Democrats for sharing misinformation on X's Community Notes. *Proc. Natl. Acad. Sci. USA* **2025**, *122*, e2502053122. [CrossRef] [PubMed]
24. Matamoros-Fernández, A.; Jude, N. The importance of centering harm in data infrastructures for 'soft moderation': X's Community Notes as a case study. *New Media Soc.* **2025**, *27*, 1986–2011. [CrossRef]
25. Hallinan, B.; Reynolds, C.; Kuperberg, Y.; Rothenstein, O. Aspirational platform governance: How creators legitimise content moderation through accusations of bias. *Internet Policy Rev.* **2025**, *14*, 1–28. [CrossRef]
26. Augenstein, I.; Bakker, M.; Chakraborty, T.; Corney, D.; Ferrara, E.; Gurevych, I.; Hale, S.; Hovy, E.; Ji, H.; Larraz, I.; et al. Community Moderation and the New Epistemology of Fact Checking on Social Media. *arXiv* **2025**, arXiv:2505.20067.
27. Boyd, D. Social network sites as networked publics: Affordances, dynamics, and implications. In *A Networked Self: Identity, Community, and Culture on Social Network Sites*; Papacharissi, Z., Ed.; Routledge: New York, NY, USA, 2010; pp. 39–58.
28. van Dijck, J. *The Culture of Connectivity: A Critical History of Social Media*; Oxford University Press: Oxford, UK, 2013. [CrossRef]
29. Bucher, T.; Helmond, A. The affordances of social media platforms. In *The SAGE Handbook of Social Media*; Burgess, J., Marwick, A., Poell, T., Eds.; SAGE Publications: London, UK, 2018; pp. 233–253.
30. Highfield, T.; Leaver, T. Instagrammatics and digital methods: Studying visual social media, from selfies and GIFs to memes and emoji. *Commun. Res. Pract.* **2016**, *2*, 47–62. [CrossRef]
31. Pew Research Center. Social Media Fact Sheet. 2025. Available online: https://www.pewresearch.org/internet/fact-sheet/social-media/ (accessed on 28 June 2026).
32. Schreier, M. *Qualitative Content Analysis in Practice*; SAGE Publications: London, UK, 2012. [CrossRef]
33. Krippendorff, K. *Content Analysis: An Introduction to Its Methodology*, 4th ed.; SAGE Publications: Thousand Oaks, CA, USA, 2019. [CrossRef]
34. Altheide, D.L. Ethnographic content analysis. *Qual. Sociol.* **1987**, *10*, 65–77. [CrossRef]
35. Franzke, A.S.; Bechmann, A.; Ess, C.M.; Zimmer, M. (Eds.) *Internet Research: Ethical Guidelines 3.0*; Association of Internet Researchers (AoIR): Chicago, IL, USA, 2020. Available online: https://aoir.org/reports/ethics3.pdf (accessed on 28 June 2026).
36. Napoli, P.M. *Social Media and the Public Interest: Media Regulation in the Disinformation Age*; Columbia University Press: New York, NY, USA, 2019.
37. Buckingham, D. *The Media Education Manifesto*; John Wiley & Sons: Cambridge, UK, 2019.
38. Livingstone, S.; Mascheroni, G.; Staksrud, E. European research on children's internet use: Assessing the past and anticipating the future. *New Media Soc.* **2018**, *20*, 1103–1122. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.