Community Notes as a Boundary Object for Distributed Oversight

Michal Shaharabani, Irit Hadar and Osnat Mokryn

Department of Information Systems, University of Haifa, Israel
Department of Information Systems, University of Haifa, Israel
Department of Information Systems, University of Haifa, Israel

Abstract

Social networks have amplified the spread of misinformation, raising persistent challenges around trust, interpretation, and governance of online information. This study examines Community Notes, X's crowd-sourced mechanism for adding context to posts, as an AI-mediated socio-technical boundary object that enables a form of distributed epistemic oversight. Rather than acting as a centralized fact-checker or automated decision system, Community Notes coordinates collective sense-making by mediating the visibility of contextual information across different groups of users. We report findings from an interview-based pilot study with six participants, analyzed through the Theory of Planned Behavior. The analysis reveals wide variation in attitudes, perceived social norms, and perceived behavioral control, shaped in part by how participants understand, and often misunderstand, the system's role and decision process. Engagement does not align cleanly with trust alone. Participants may engage despite distrust, disengage despite trust, or participate strategically to counter perceived bias. These findings suggest that Community Notes functions less as a truth-adjudication mechanism and more as a boundary object that structures oversight through contextual framing and signaling. The results inform the design of a large-scale survey aimed at testing the generalizability of these observations.

Keywords

Community Notes, Human oversight of AI systems, Boundary objects, Epistemic oversight

1. Introduction

The proliferation of misinformation on social media poses a serious threat to democratic discourse, public health, and societal trust [1]. Platforms such as Facebook, X (formerly Twitter), and YouTube have amplified the reach of misleading content, which has been shown to spread farther, faster, and deeper than truthful information, driven in part by its novelty and emotional resonance [1]. Researchers have highlighted not only the vast amount of misinformation online, but also the difficulty users face in consistently identifying credible information [2].

Interventions to address misinformation have included algorithmic detection (e.g., identifying coordinated inauthentic activity), third-party fact-checking, and crowdsourced verification [3, 4]. Crowdsourced systems leverage the "wisdom of the crowd" to collectively assess and contextualize information, offering scalable solutions that complement expert-based approaches [5, 6].

Among these, the *Community Notes* system implemented by X enables users to collaboratively contribute contextual information to potentially misleading posts. This model prioritizes *diverse agreement*, that is, consensus among contributors from different ideological backgrounds, as a condition for public visibility of notes [7, 6], aligning with research suggesting bipartisan consensus enhances perceived trustworthiness [8]. While Community Notes relies on human contributions, decisions about which notes become visible are governed by algorithmic processes, making it an AI-mediated system that coordinates human judgment at scale.

In this work, we treat Community Notes not only as a feature or moderation tool, but as an AI-mediated object system composed of three tightly coupled elements: user-generated contextual notes, algorithmic mechanisms that determine note visibility, and social interpretations of credibility and

Joint Proceedings of the ACM Intelligent User Interfaces (IUI) Workshops 2026, March 23-26, 2026, Paphos, Cyprus
*Corresponding author.
thepicachu@gmail.com (M. Shaharabani); hadari@is.haifa.ac.il (I. Hadar); omokryn@is.haifa.ac.il (O. Mokryn)
0000-0002-9421-8566 (O. Mokryn)

intent. This object system does not produce authoritative judgments, but it shapes how contested information is interpreted by signaling when and how additional context should be considered. As such, it enables a form of oversight that is distributed across contributors, mediated by algorithmic rules, and enacted through user attention rather than enforcement.

Despite the platform's transparency and reliance on diverse perspectives, it remains unclear whether users understand Community Notes mechanisms or are motivated to engage with them as an AI-mediated oversight system. Guided by the Theory of Planned Behavior (TPB) [9], this study examines how three key factors: (i) users' attitudes toward Community Notes, (ii) perceived social norms regarding its use, and (iii) perceived behavioral control in contributing to or evaluating notes, influence both contributors' and general users' intentions to rely on or participate in the system. Specifically, we ask: *How do attitudes, subjective norms, and perceived behavioral control shape user engagement with Community Notes in X?*

2. Theoretical Background

The spread of misinformation online has been extensively documented, with studies showing that false information can travel farther, faster, and deeper than truthful content due to its novelty and emotional appeal [1]. Coordinated inauthentic behavior and algorithmic filtering further exacerbate this spread by reinforcing echo chambers [3, 10]. These challenges have driven interest in scalable, trustworthy verification mechanisms.

Crowdsourced verification leverages the collective judgment of users to assess content accuracy. Prior work has shown that, under certain conditions, aggregated crowd ratings can match or exceed expert fact-checkers in accuracy [6]. Trust in community-based fact-checking systems is influenced by the transparency of the process, the credibility of cited sources, and mechanisms that encourage diverse participation and fair evaluation of contributions [11]. X's *Community Notes* is a unique implementation that requires *diverse agreement*, i.e., consensus from contributors with differing ideological perspectives, before a note becomes public [7].

Although Community Notes does not rely on automated content generation or classification, it is nonetheless an AI-mediated system in the sense that algorithmic processes operationalize human judgments at scale. Decisions about which notes become visible are governed by computational models that infer cross-ideological agreement and contributor reliability. These models do not determine truth, but they shape epistemic outcomes by controlling visibility and timing. Oversight in this context therefore targets not an autonomous AI agent, but a socio-technical system in which algorithmic mediation plays a central coordinating role.

Early empirical analyses highlight both strengths and challenges of the system. While notes often fill important contextual gaps, many fail to appear due to insufficient cross-ideological ratings, raising questions about the balance between inclusivity and timeliness [12, 13]. Moreover, contributors' ideological self-identification only partially predicts their rating behavior, suggesting that diversity-based models may have complex, unintended dynamics.

Beyond system-specific factors, user engagement can be understood through the *TPB* [9]. TPB posits that behavioral intentions are shaped by three core components:

- **Attitude toward the behavior** - the individual's overall evaluation of the behavior as favorable or unfavorable, based on beliefs about its outcomes and their desirability. In the context of this study, it refers to whether participants view contributing to or relying on Community Notes as worthwhile, accurate, and beneficial, or alternatively as ineffective, biased, or open to manipulation.
- **Subjective norms** - the perceived social pressure from others to perform or not perform the behavior, reflecting expectations and norms within one's social context. In the case of Community Notes, this includes whether participants feel that peers, online communities, or broader societal groups expect them to contribute or value their participation, and whether they believe that others contribute with integrity.

- **Perceived behavioral control** - the perceived ease or difficulty of performing the behavior, influenced by past experiences, knowledge, and available resources, and closely related to self-efficacy. Here, it reflects whether participants feel knowledgeable and capable of contributing effectively to Community Notes, understanding the rating process, and believing that their contributions can influence which notes are published.

In this framework, contributors' decisions to participate and users' acceptance of Community Notes as credible are shaped not only by beliefs about the system's accuracy but also by social expectations and the degree to which individuals feel capable of making a meaningful impact.

From an oversight perspective, attitudes, norms, and perceived behavioral control do not only shape willingness to engage, but also shape users' perceived capacity to exercise meaningful oversight. Misunderstandings of system mechanics may limit users' ability to evaluate when notes are trustworthy, to decide whether intervention is warranted, or to assess whether participation can have impact. TPB therefore provides a lens for examining how distributed oversight emerges, weakens, or fails in practice.

In sum, while prior research highlights the potential of diverse crowdsourced models to combat misinformation [12, 11, 13], little is known about the cognitive factors that shape the intention to engage in such systems. Addressing this gap is the central motivation of the present study, while recruiting the TPB as the theoretical framework for a comprehensive analysis. While several adoption-oriented frameworks, such as the Technology Acceptance Model (TAM; 14) and the Unified Theory of Acceptance and Use of Technology (UTAUT; 15), have been widely used to explain technology acceptance and system use, they primarily emphasize instrumental determinants such as perceived usefulness, ease of use, and facilitating conditions. In contrast, TPB 9 focuses on volitional behavior by explicitly modeling behavioral intentions as a function of attitudes, subjective norms, and perceived behavioral control. This emphasis makes TPB particularly suitable for examining voluntary and normatively embedded forms of participation in socio-technical oversight systems. Accordingly, TPB is well aligned with the study of Community Notes, where engagement is not mandated or efficiency-driven, but socially situated and shaped by users' beliefs, perceived norms, and sense of agency.

3. Methodology

In this pilot study, TPB was used as a unifying framework to examine two related but distinct forms of engagement: contribution (writing and rating notes) and reliance (reading and interpreting notes). While these behaviors differ, both involve voluntary, belief-driven interaction with the system. Accordingly, this research adopts a mixed-methods approach in two stages: (1) a qualitative exploratory stage based on semi-structured interviews, and (2) a planned quantitative stage involving a large-scale survey

1. **Qualitative stage (completed)** - Semi-structured interviews were conducted with six X users: two registered Community Notes contributors and four non-contributors. The interviews aimed to elicit participants' attitudes toward Community Notes, perceived social norms surrounding its use, and perceived behavioral control in relation to contributing to or relying on the system. In addition, participants were asked to describe their understanding of Community Notes' purpose, decision logic, and algorithmically mediated visibility mechanisms.
   Prior to data collection, the interview guide was piloted in several preliminary interviews to refine question clarity, sequencing, and conceptual focus. Feedback from these pilots informed adjustments to wording and structure, ensuring that questions were accessible to participants with varying levels of familiarity with Community Notes and social media systems more broadly. Participants differed in their level of experience with the system, ranging from active contributors with hands-on experience in writing and rating notes to general users whose engagement was limited to exposure and interpretation. This variation enabled exploration of how differing levels of familiarity and perceived competence shape engagement and oversight-related perceptions.
2. **Quantitative stage (planned)** - A large-scale survey grounded in the TPB will be designed and distributed to relevant users. The survey will enable measuring attitudes, subjective norms, and

perceived behavioral control in the context of participating in Community Notes, along with intentions to rely on and engage with the fact-checking system. When designing the survey we will translate key qualitative themes into survey items. This approach ensures that the constructs included in the survey will be grounded in real participant experiences, enabling a more precise examination of the factors influencing engagement in the Community Notes system.

Participants varied in age (40s-70s), gender, and professional background, and included both active Community Notes contributors and general X users. All participants had prior exposure to Community Notes, though their levels of engagement and understanding of the system varied. Semi-structured interviews focused on participants' perceptions of Community Notes' purpose, trustworthiness, engagement motivations, and understanding of the rating and visibility mechanisms. This paper reports the preliminary findings from the qualitative stage.

4. Preliminary Results

The qualitative study included six participants, recruited via purposive sampling to capture variation in role and experience. Two participants were registered Community Notes contributors, and four were general X users (non-contributors). Participants (4 men, 2 women) ranged in age from their 40s to early 70s and came from diverse professional backgrounds, including a lawyer, an accountant, a high-tech professional, a security officer, a journalist and a homemaker. All participants were active X users with prior exposure to Community Notes, although levels of engagement and understanding varied substantially. Contributors had formal access to writing and rating notes, while non-contributors interacted with the system primarily through exposure and reliance.

The pilot stage of the research revealed patterns consistent with the TPB: attitudes toward the behavior, subjective norms, and perceived behavioral control. While the small sample limits systematic comparison, differences emerged in how contributors and non-contributors described their engagement. Contributors more frequently articulated active and sometimes strategic participation, including a willingness to invest effort in sourcing evidence and influencing outcomes despite uncertainty or distrust. In contrast, non-contributors described lower perceived agency, encountering Community Notes opportunistically and expressing uncertainty about whether individual actions such as reading, reacting to, or ignoring notes had any meaningful impact.

Attitudes toward the behavior: Participants expressed a wide spectrum of views. Some demonstrated strong trust in the system, believing it to be accurate because every correction they had seen was aligned with their understanding of the facts. For example, one participant noted, "100% of the corrections I saw were accurate". Others expressed deep skepticism, viewing the system as biased or easily manipulated. For example, "99% of the people posted it just to be controversial", or "Planting bots that can influence the rating". In some cases, distrust motivated participation. For example, rating in order to counter perceived bias, while in others, trust stemmed from personal experience rather than from knowledge of the process. In general, participants engaged selectively, reading or rating only topics of personal interest.

Subjective norms: Social influence varied considerably. In some cases, direct pressure from politically affiliated groups led participants to join and rate. For example, "It popped up in one of the protest groups, and then they asked everyone to become Community Notes contributors". Others were drawn in by broader online events, such as "There was some commotion around a certain topic, people talked about it". In some cases, influence came from prominent online figures, as one participant described, "There's a tweeter named [omitted[1]] who brought evidence". Participants also expressed assumptions about how others behave; one remarked they behave "exactly like me", while others rejected such assumptions; for example, one participant answered "Absolutely not" to the question "Do you rely on others to act like you?" In most cases, discussions of Community Notes with friends or colleagues were rare, suggesting that direct, ongoing social pressure to engage is limited.

[1]We removed the X handle to safeguard personally identifiable information (PII).

Perceived behavioral control: Confidence in one's ability to contribute effectively varied. Some participants rated notes confidently without consulting official guidelines, acting on the assumption that their approach was valid. Others learned the mechanics gradually and reported increased confidence over time. Barriers to contribution included a lack of time, uncertainty over whether contributions influence publication, and confusion about key concepts. For example, one participant noted, "I haven't come across any Community Notes for a long time", indicating that limited exposure to the feature reduced their opportunities and perceived ability to engage. Some reported a lack of understanding or interest; for example, "I have no idea, they explained it to me several times. It didn't really interest me". Time constraints were also mentioned: "Maybe if I had a bit more time", as well as general distrust in the system's safeguards. Certain participants, while acknowledging their potential to make a significant contribution, refrained from doing so because, as one explained, "I'm not entirely sure how it works", and framing participation as a personal choice: "It comes down to a personal choice whether to read and respond". Interestingly, none of the participants demonstrated a fully accurate understanding of how the Community Notes system operates. Several believed they understood it, but their explanations contained substantial misconceptions, including among active contributors. This pattern suggests that oversight may be exercised without a shared understanding of how the system operationalizes agreement. Additionally, confidence without procedural understanding raises questions about the reliability of distributed oversight.

Overall, the findings indicate that attitudes toward Community Notes, perceptions of social expectations, and perceived behavioral control interact in complex ways to shape engagement. They reveal that participation can occur despite distrust, that trust does not always translate into broad engagement, and that both accurate and inaccurate beliefs about the system - as well as assumptions about how others behave - influence behavioral intentions.

5. Discussion and Next Steps

The findings of this pilot study suggest that Community Notes operates as a form of soft, distributed oversight rather than as a corrective or adjudicative mechanism. Oversight in this system is not exercised through content removal or authoritative judgments, but through contextual framing and visibility signals that shape how information is interpreted by readers. Contributors monitor content and provide context, while algorithmic mechanisms coordinate which interventions become visible. In this sense, oversight emerges from the interaction between human participation and algorithmic mediation, rather than from either component alone.

Viewed as a boundary object, Community Notes accommodates multiple and sometimes conflicting interpretations of its purpose and legitimacy. Some participants treat it as a corrective system aimed at improving information quality, others view it as a political arena, and some see it as unreliable or symbolic. These interpretations coexist without being resolved by the system. Our findings show that this flexibility enables participation but also produces uneven forms of oversight. Participants often act without a shared understanding of system mechanics, and confidence in contributing does not necessarily reflect procedural knowledge. As a result, oversight may be exercised inconsistently, shaped by personal beliefs, assumptions about others' behavior, and misconceptions about how agreement and visibility are operationalized.

From the perspective of the TPB, attitudes, subjective norms, and perceived behavioral control shape not only engagement, but also perceived capacity to exercise meaningful oversight. Distrust does not preclude participation and may even motivate it, particularly when contributors aim to counter perceived bias or influence outcomes. At the same time, trust and understanding do not guarantee engagement when perceived control is limited. These patterns suggest that distributed oversight depends not only on willingness to participate, but also on shared understanding of system roles and limits. For AI-mediated systems that rely on human input to coordinate oversight, this highlights the importance of oversight literacy and transparency, alongside mechanisms that support participation. TPB was used as an analytic lens for interpretation rather than as a coding scheme, allowing contradictory accounts to coexist while

still informing construct development for the subsequent survey. This study has several limitations, including the small and relatively homogeneous sample, reliance on self-reported perceptions, and focus on a single platform. The planned large-scale survey will enable systematic testing of these observations across a broader population and allow examination of how different configurations of attitudes, norms, and perceived control relate to engagement and oversight outcomes. More broadly, this work contributes to ongoing discussions about human oversight of AI-mediated systems by showing how oversight can be enacted through distributed, interpretive practices rather than centralized control, and how user perceptions play a critical role in shaping whether such systems function as intended.

Declaration on Generative AI

Generative AI tools were used only as a language support tool and did not influence the scientific content, analysis, or interpretation of the research.

References

[1] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, Science 359 (2018) 1146–1151.

[2] G. Pennycook, A. Bear, E. T. Collins, D. G. Rand, The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings, Management Science 66 (2020) 4944–4957. doi:10.1287/mnsc.2019.3478.

[3] E. Ferrara, Disinformation and social bots, First Monday 22 (2017). URL: https://firstmonday.org/ojs/index.php/fm/article/view/7090. doi:10.5210/fm.v22i11.7090.

[4] K. Shu, A. Sliva, S. Wang, J. Tang, H. Liu, Fake news detection on social media: A data mining perspective, ACM SIGKDD Explorations Newsletter 19 (2017) 22–36.

[5] J. Allen, A. Arechar, G. Pennycook, D. Rand, Evaluating the quality and impact of community-based fact-checking on twitter, Harvard Kennedy School Misinformation Review (2022).

[6] C. Martel, D. G. Rand, Fact-checker warning labels are effective even for those who distrust fact-checkers, Nature Human Behaviour 8 (2024) 217–229. doi:10.1038/s41562-024-01973-x.

[7] M. L. Gordon, et al., Jury learning: Integrating dissenting voices into machine learning models, in: Proceedings of the 2022 AAAI/ACM Conference on AI Ethics and Society (AIES '22), 2022, pp. 654–665.

[8] G. Pennycook, D. G. Rand, The psychology of (fake) news: Accepting, sharing, and correcting misinformation, Nature Reviews Psychology 1 (2022) 46–57. doi:10.1038/s44159-021-00006-y.

[9] I. Ajzen, The theory of planned behavior, Organizational Behavior and Human Decision Processes 50 (1991) 179–211. doi:10.1016/0749-5978(91)90020-T.

[10] E. Bakshy, S. Messing, L. A. Adamic, Exposure to ideologically diverse news and opinion on facebook, Science 348 (2015) 1130–1132. doi:10.1126/science.aaa1160.

[11] N. Proollochs, Community-based fact-checking on twitter's birdwatch platform, in: Proceedings of the Sixteenth International AAAI Conference on Web and Social Media (ICWSM 2022), 2022, pp. 794–804.

[12] D. Wirtschafter, M. Majumder, Future challenges for online, crowdsourced content moderation, Journal of Online Trust and Safety 5 (2023) 23–45. URL: https://onlinetrustsafetyjournal.org/article/view/345. doi:10.1234/jots.v5i1.345.

[13] T. Renault, D. Restrepo-Amariles, A. Troussel-Clément, Collaboratively adding context to social media posts reduces the sharing of false news, arXiv preprint arXiv:2404.02803v1 (2024). Université Paris 1 Panthéon-Sorbonne and HEC Paris.

[14] F. D. Davis, Perceived usefulness, perceived ease of use, and user acceptance of information technology, MIS Quarterly 13 (1989) 319–340. doi:10.2307/249008.

[15] V. Venkatesh, M. G. Morris, G. B. Davis, F. D. Davis, User acceptance of information technology: Toward a unified view, MIS Quarterly 27 (2003) 425–478. doi:10.2307/30036540.