Community Moderation and the New Epistemology of Fact Checking on Social Media

Isabelle Augenstein[1], Michiel Bakker[2], Tanmoy Chakraborty[3,*], David Corney[4], Emilio Ferrara[5], Iryna Gurevych[6], Scott Hale[7], Eduard Hovy[8], Heng Ji[9], Irene Larraz[10], Filippo Menczer[11], Preslav Nakov[12], Paolo Papotti[13], Dhruv Sahnan[12], Greta Warren[1], and Giovanni Zagni[14]

[1]University of Copenhagen, Nørregade 10, 1172 København, Denmark
[2]Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA, United States of America
[3]Indian Institute of Technology Delhi, New Delhi, 110016, India
[4]Full Fact, 17 Oval Way, London, SE11 5RR, United Kingdom
[5]University of Southern California, Los Angeles, CA 90007, USA
[6]Technical University of Darmstadt, Hochschulstraße 10, D-64289 Darmstadt, Germany
[7]University of Oxford, Broad St, Oxford OX1 3AZ, United Kingdom
[8]The University of Melbourne, Grattan Street, Parkville, Victoria 3010, Australia
[9]University of Illinois Urbana-Champaign, 506 S. Wright St. Urbana, IL 61801-3633, USA
[10]Newtrales, C/Vandergoten 1, 28014 Madrid, Spain
[11]Indiana University, 1015 E 11th St., Bloomington, IN 47408, USA
[12]Mohamed bin Zayed University of Artificial Intelligence, Masdar City, Abu Dhabi, 7909, United Arab Emirates
[13]EURECOM, Campus SophiaTech, 450 Route des Chappes, CS 50193 - 06904 Biot, FRANCE
[14]Pagella Politica/Facta, viale Monza 259/265, Milano, 20125, Italy
[*]Corresponding author, Email: tanchak@iitd.ac.in

ABSTRACT

Social media platforms have traditionally relied on internal moderation teams and partnerships with independent fact-checking organizations to identify and flag misleading content. Recently, however, platforms including X (formerly Twitter) and Meta have shifted towards community-driven content moderation by launching their own versions of crowd-sourced fact-checking – Community Notes. If effectively scaled and governed, such crowd-checking initiatives have the potential to combat misinformation with increased scale and speed as successfully as community-driven efforts once did with spam. Nevertheless, general content moderation, especially for misinformation, is inherently more complex. Public perceptions of truth are often shaped by personal biases, political leanings, and cultural contexts, complicating consensus on what constitutes misleading content. This suggests that community efforts, while valuable, cannot replace the indispensable role of professional fact-checkers. Here we systemically examine the current approaches to misinformation detection across major platforms, explore the emerging role of community-driven moderation, and critically evaluate both the promises and challenges of crowd-checking at scale.

Introduction

Social media platforms empower users to share opinions and perspectives at scale. This openness brings the persistent challenge of dealing with harmful, misleading, or otherwise objectionable content without unduly constraining freedom of expression. To navigate this tension, platforms implement content moderation policies aimed at protecting users from potentially dangerous material while preserving the integrity of public discourse. The responsibility of enforcing these rules traditionally falls upon teams of experts such as content moderators, supplemented by veracity judgments provided by third-party independent fact-checkers[1–3] as well as by automatic Artificial Intelligence (AI) systems working in tandem[4–6]. In practice, these platform *enforcers* determine whether to remove content, restrict its visibility, or attach warning disclaimers to a post, depending on its potential to cause harm. However, both AI-based and centralized manual moderation have limitations. AI tools for content moderation, while scalable, are marred by high rates of *false positives*, often over-flagging benign content and unfairly reprimanding users, and *false negatives*, letting truly harmful content slip by undetected[7–9]. Recent advances in large language models (LLMs) have shown some promise in automatic veracity prediction[10,11] to support fact-checkers but suffer from issues with factuality[12] and utility in practice[13,14]. Fact-checking experts are reliable, but cannot keep up with the relentless pace of user-generated content posted online[15,16].

The sheer volume and velocity of online information has strained traditional fact-checking models, leading to innovations in crowd-sourced moderation approaches such as Community Notes[17], which aim to leverage crowd wisdom for broader and faster coverage, albeit with its own set of considerations regarding quality and bias. In an effort to keep content moderation as democratic as possible, the core idea behind this approach is to leverage the collective input of users on the platform (i.e., *community*) to add context to posts that may violate content policies or contain misleading information. In this approach, platforms maintain automatic systems and internal teams in place to remove illegal or severely harmful posts containing harassment, violence, sexual exploitation, and drug-related content, while the remainder of posts are subject to community-driven moderation. In January 2025, Meta revealed a major policy change, announcing that it would end the use of third-party fact-checkers on its platforms due to alleged bias[18]. Meta stated that content that many users might view as acceptable political commentary — albeit to be consumed with a pinch of salt — may have been unnecessarily suppressed. This argument is disputed by fact-checkers, who point out that they take serious steps to maintain political impartiality, including regular internal and external reviews[19].

Neither professional fact-checking alone nor nascent community moderation systems like Community Notes offer a perfect solution to online misinformation. While professional fact-checking offers depth and rigor, its scalability is limited. Conversely, community-driven models like Community Notes promise scalability and diverse perspectives but must navigate challenges of consensus-building and potential manipulation. Additionally, it builds upon an epistemological proposition that facts are subject to consensus and negotiation, rather than objective or indisputable as traditionally intended by professional fact-checkers. Such a shift has far-reaching potential consequences for the global information eco-system more broadly. This paper explores how these distinct approaches have been deployed in content moderation, critically assesses the promises of community moderation, and highlights how collaboration between communities, experts, and technical innovations can address pervasive online misinformation.

Community-Driven Content Moderation

The concepts underpinning community-driven content moderation have a long history[20]. Wikipedia, for instance, has long operated on the principle of collaborative governance, relying on its users to faithfully curate and manage information for the platform[21]. Similarly, Reddit delegates some autonomy to its users, allowing a subset of users to moderate content within their respective subreddits[22]. Many social media platforms lean on user reports, not only to flag harmful content that evades content moderators, but also to help calibrate policies in response to concerns about certain types of content[18, 23]. However, user reports that inform content moderation decisions implemented by the platforms differ from more structured approaches to community fact-checking.

In the domain of community-driven content moderation on social media platforms, Twitter (now X) launched *Birdwatch*[17] in early 2021, the first large-scale initiative in this space, and later rebranded it *Community Notes*[24, 25]. While the program's initial iteration exposed significant shortcomings—such as a vulnerability towards targeted manipulation attempts and partisan bias affecting the notes' writing style and approval, the company has since invested a substantial amount of resources into its refinement[26, 27]. Key improvements include a more sophisticated algorithm to calculate helpfulness of notes, which rewards notes endorsed by a diverse set of users rather than a simple majority; and eligibility criteria for contributors to ensure participation by genuine users[25]. Following X's lead, other social media giants, such as Meta and, to a lesser extent, TikTok and Weibo, have recently pivoted in favor of a similar community-driven moderation approach over hired experts for moderation on their platforms[28–30]. Some of these platforms posit community notes as a one-size-fits-all solution to the limitations of fact-checker-led content moderation. Community notes have indeed shown encouraging results on several fronts where moderation assisted by third party fact-checkers is limited, such as content coverage[31–33]. However, most of these results and the portrayal of community-driven content moderation as the definitive solution rest on overly-optimistic assumptions about the integrity, diversity, and efficiency of user collaboration. Just as citizen journalism's initial promise to democratize information has faltered and been critiqued for its issues with capacity, reliability and lack of professional standards[34, 35], similar limitations may manifest in community notes.

In this paper, we review the current implementation of community notes. As illustrated in Figure 1, the frameworks deployed by platforms such as X, Meta, and TikTok follow a similar multi-step process[3, 25, 30, 36, 37]. All new posts first go through a pre-moderation step, where an automated "harmfulness" classifier assesses the harm potential of the content. Based on the nature of the content and the inferred type of harm, the content is categorized as either *restricted* or *less harmful*.

Posts with content that poses severe harm, such as calls for violence or terrorism, depictions of child sexual exploitation, and drug-related activities, are deemed to be completely unacceptable by most platforms. Moreover, the spread of these posts also exposes platforms to expensive lawsuits for failing to protect their users from digital harassment[38, 39]. Such content is labeled as *restricted* at the pre-moderation step itself and is not published on the platforms. Additionally, the platforms monitor user reports to identify and remove content violations that seep through the automated system. Similarly, content about which a classifier is uncertain may go directly to a queue for a manual check[9]. Notably, X took down over 3 million posts from public view in the latter half of 2024, either via automated flagging systems or human review[40].

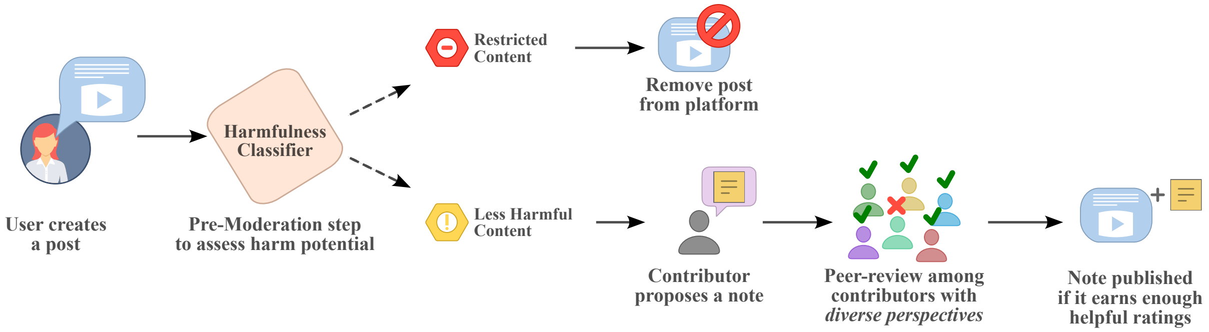
User creates a post
Harmfulness Classifier
Pre-Moderation step to assess harm potential
Restricted Content
Remove post from platform
Less Harmful Content
Contributor proposes a note
Peer-review among contributors with *diverse perspectives*
Note published if it earns enough helpful ratings

Figure 1. An overview of the community-driven content moderation framework as proposed by major social media platforms like X, Meta, and TikTok. The moderation process is divided into two main stages: (i) **Pre-Moderation** using AI classifiers, which categorize content as either *restricted* — blocked from appearing on the platform — or *less harmful*, which is passed on for community moderation; and (ii) **Community Moderation**, where eligible volunteers may propose additional context that undergoes peer review by other contributors with diverse perspectives before being published after a consensus is achieved.

Posts that are not automatically restricted may still contain misleading or harmful content. Such posts are subject to community-driven content moderation, where eligible users may propose *community notes* providing context that clarifies why the content may be wrong or misleading. The proposed note, then, undergoes a peer-review process, where other community note contributors with diverse perspectives rate its *helpfulness*. For a note to appear publicly, it must earn enough helpful ratings that pass the platform's acceptability threshold. Once rated as *helpful*, the note is displayed alongside the original post. While Meta previously prevented posts labeled by independent fact-checkers as false or misleading from being algorithmically promoted, this is not currently the case for posts with community notes. However, it is important to note that what constitutes 'diverse perspectives' remains unclear, as platforms do not clearly define diversity itself – only how it is quantified in the algorithm using historical mutual disagreement between users on the perceived helpfulness of other community notes.

The promises of community moderation

Social media platforms claim that community moderation offers several improvements over expert fact-checking:

- **The Community Notes model democratizes content moderation.** Conventionally, fact-checkers advise platforms on potentially harmful mis/disinformation, prompting actions like labeling, limiting spread, or removal. Meta claims this has led to user dissent over perceived censorship[18]. In contrast, community moderation allows all (non-severely harmful) posts to remain public, letting users decide if a *note of caution* is needed. This may reduce perceived bias and encourage pluralistic interpretations. However, Wikipedia shows that such systems can be undermined by dominant editors or collusive groups[41]. Community moderation also risks favoring popularity over truth. Moreover, due to echo chambers on social media[42,43], community notes may not reach ideologically diverse users, weakening their effectiveness[44].
- **Community moderation can scale up and speed up misinformation detection.** The volume of user-generated content far exceeds what professional fact-checkers can handle, forcing them to prioritize the most harmful and verifiable claims. Payment models tied to fact-check volume can incentivize checking easier claims, leading to biases and leaving harmful content unchecked. Still, this *selection bias* is arguably justified, as it targets the most dangerous content. Community notes, by contrast, allow a broader pool of volunteers to verify more diverse content. Yet, self-selection remains a challenge – few have the time or motivation to participate, as seen in the skewed demographics of Wikimedia editors[45]. Discovery is another issue: due to personalization and polarization, knowledgeable users may not encounter misleading posts[43]. Moreover, expert verification often comes too late – most of the engagement typically occurs before a claim is fact-checked[46]. Paid, crowd-sourced moderation can be faster and scale better[44], but has outperformed experts only in limited cases. Many useful notes remain unpublished due to contributor disagreements, and it is unclear if such systems cover more content overall[31,47].
- **Community moderation is less intrusive.** Community-driven moderation shows peer-approved notes non-intrusively alongside posts[18,25], allowing users to engage with or ignore them. Earlier expert-led systems often used more intrusive warnings, requiring user action to proceed. Community notes usually avoid verdicts (e.g., true/false), instead offering missing context – useful for ambiguous claims or grey areas like dogwhistles[48], and less likely to polarize. However, studies show that expert warning labels effectively reduce belief in and spread of misinformation, even among skeptics[49,50]. Ultimately, the 'intrusiveness' of moderation is a design choice, and fact-checker systems could adopt similar display styles. Therefore, a direct comparison of the effectiveness of both approaches remains necessary.

Community Notes versus Third-Party Fact-Checking

The rhetoric and policy shifts by social media platforms such as Meta and X suggest that community notes are a salve to the issues with fact-checking[18, 24]. We believe this is a false dichotomy: the two approaches are deeply intertwined[51]. Here, we present a balanced investigation of the extent to which community notes delivers on the claimed benefits. We shed light on potential risks for users and platform integrity, which make community notes ill-suited as a comprehensive replacement of moderation experts. Table 1 presents a brief overview of our discussions in subsequent subsections, offering a comparative breakdown, across various indicators, of the two moderation approaches: use of third-party fact-checkers and reliance on community moderation.

Are community notes more scalable?

By harnessing the "wisdom of the crowd," the community notes model has a strong potential to enhance the scalability and quicken the response times of content moderation on social media platforms[44]. However, analyses of similar crowd-sourced approaches demonstrate competing evidence that cast doubt on the claims of increased effectiveness over expert-led moderation.

- **Volume:** Community notes can be written and voted on by any platform user (subject to a minimum quality check[52]), while professional fact-checkers are typically limited to individuals with training in journalism. Given the limited capacity of fact-checking projects (as of May 2024, there were 439 independent professional fact-checking projects in 111 countries[53]), community notes hold the potential to address a much larger volume of misleading claims than fact-checkers alone. However, given that just a small proportion of proposed notes reach publication status[31], the volume of moderated posts does also not correspond to the (output) volume of warning labels.
- **Breadth:** Fact-checkers outside Western, Educated, Industrialised, Rich, and Democratic (WEIRD) nations face challenges like limited press freedom[54], data scarcity[55], financial constraints[14, 56], and even physical threats[57], despite support from global networks[58]. Community notes can extend coverage to regions where professional fact-checking is constrained. Pseudonymity allows laypeople and citizen journalists to flag misinformation in risky environments. Professionals prioritize high-virality or high-harm claims[59, 60], whereas community notes can address a wider range—provided volunteers are present. However, contributor demographics (e.g., Wikipedia) remain skewed toward Western males[45, 61]. Only 20% of contributors have written notes that reach consensus[62], indicating that self-selection may limit both participation and topic diversity.
- **Expertise:** Fact-checking is a skilled and complex task that requires specific experience, expertise, and data access, for example, in identifying patterns of misinformation and uncovering large-scale disinformation campaigns. Furthermore, verifying complex or high-stakes claims (e.g., relating to health, science, or economics) often requires specific expertise, which most community note-writers are unlikely to possess. Often, the information needed to verify a claim is not available online and requires creating new knowledge, e.g., by directly contacting experts or first-hand witnesses[14, 63]. In these cases, crowd-sourced fact-checking tends to rely on existing analyses by professional fact-checkers[51, 64]. Community note writers often target lower-risk misleading posts, such as scams[51] and claims that have previously been fact-checked[31].
- **Speed:** While professional fact-checking articles undergo lengthy cross-checks and editorial oversight to ensure quality, community notes have been posited as a method of expediting the fact-checking process. However, there is mixed evidence in support of this claim: one study found 20–30 ideologically diverse laypeople can reach accurate verdicts faster than experts[44], but another showed the 'crowd' was quicker than experts in less than 6% of cases[31]. A key bottleneck is the requirement for cross-ideological agreement before publishing a note, leading to an average delay of 15.5 hours[33] and many helpful notes never being published[47]. The system may also be vulnerable to manipulation by coordinated actors suppressing valid notes[65]. In collaborative models like Wikipedia-style editing, laypeople were often faster than experts only because they could reference existing professional fact-checks[64], indicating the complementarity between expert and volunteer fact-checking.

Are community notes more trustworthy?

One of the areas in which community notes hold the greatest promise is their potential to increase public trust in fact-checking, content moderation, and social media platforms. However, there are several significant barriers that must be addressed to realize this potential.

- **Democratic:** The current implementation of Community Notes on X stipulates that in order to be published, a note must receive enough 'helpful' votes from users who generally tend to disagree with one another in their votes. This method rewards writing notes that readers with distinct perspectives can agree upon, and ensures that notes are not monopolized by certain ideologies. Notes that refer to *trustworthy*[66] or *unbiased*[67] sources tend to result in higher 'helpfulness' ratings. However, due to political polarization[68], community notes on contentious political issues rarely reach a consensus[31, 47]. This approach can therefore quash fact-checks on politically-sensitive claims, even if there is a clear and indisputable

Dimension	Third-Party Fact-Checker Based Moderation	Community Notes	Our Critique
Volume	Low → Can verify only a small fraction of all content.	High → Potential to address a larger volume with thousands of volunteers	→ How well community moderation scales in practice is yet to be proven: only a small proportion of submitted community notes reach publication
Breadth	Low → Forced to focus on high-visibility claims given volume of content. → Limited in the scope of issues that can be addressed due to several factors.	High → Can cover a broader set of topics beyond expert capacity. → Anonymity can allow crowds to address sensitive claims that experts may be at risk for doing so.	→ Effectiveness is contingent on volunteers willing to engage with claims falling outside of mainstream discourse.
Expertise	High → Fact-checkers are trained on the task and can effectively address all sorts of claims.	Low → Can re-purpose existing fact-checks to address known misinformation or lower-risk claims.	→ Creation of new knowledge is often required: a skill fact-checkers are trained for, but the crowd is unlikely to exhibit.
Speed	Slow → Rigorous time-intensive quality checks to ensure factual correctness limits the number of claims verified per day.	Highly Variable → Can be much faster at the verification process by leveraging the "wisdom of the crowds".	→ In practice, community notes are faster than fact checks <10% of the time[31]. → Consensus among ideologically diverse users is a critical factor: some notes are approved quickly; other valid ones face delays or are left unpublished.
Democratic Nature	Partly → Unilateral decisions by dedicated teams and the platforms to add warning labels / remove content. → Fact-checker decisions also undergo internal peer-review by fellow fact-checkers.	Yes → The community decides if a post needs extra context and what that extra context should be. → Democratic voting process: any user can propose a note, which is made public only if a consensus on its helpfulness is reached.	→ Unclear whether a ’democratic’ process contributes to the discovery of the objective truth; may lead to valid notes remaining unpublished
Bias	Low → Forced to prioritize high-risk claims; introducing a bias in choosing the claims they can verify. → Selection bias is towards countering harmful content; no evidence of this bias resulting in suppression of authentic narratives barring recent allegations by social media platforms.	Moderate → No forced prioritization of high-visibility / high-risk claims. → Consensus among users with opposing views is required, making it inherently less biased. → Crowds are biased towards what they consider interesting to flag and whether the content is even seen by users inclined to flag it.	→ Majority of sources cited as evidence by the crowds are left-leaning in practice. → Crowds are also susceptible to various cognitive biases, while fact-checkers are trained to view content through a neutral lens.
Transparency	Moderate → Platforms offer limited justification for censorship of content, thereby making it harder to scrutinize expert decisions. → Third-party fact-checkers are bound by strict principles of transparency of bias, sources and funding	Moderate → Deliberation process can be fully transparent, with users being able to view all proposed notes and the level of agreement within the community on these notes.	→ The algorithm lacks key details, especially how diversity among users is defined. → Unclear how users may scrutinize the community's decisions. → Lack of transparency regarding bias and conflicts of interest of contributors → No onus on the crowd to outline the verification process; fact-checkers generally present this as part of their analysis of the claim.
Coordinated Adversarial Attacks	Safe → No easy way for malicious actors to attack the sanctity of the professional fact-checking process.	Vulnerable → Susceptible to coordinated attacks that challenge the credibility of authentic information, or suppress helpful notes due to dissenting views.	→ Ample evidence of malicious users colluding to inorganically boost content on social media. → Wikipedia exhibits vulnerability to collusion in the voting system for approving specific edits on the website.
Impact on misinformation beliefs and spread	High → Fact-checker labels are effective in preventing misinformation, even among users skeptical of fact-checkers.	Mixed → Found effective in preventing spread of misinformation in some cases. → Evidence suggests it expedites voluntary retraction of misleading posts. → Considerable time delay between publication of the post and the community note.	→ Notes are not "promoted" by the platform in any way. People who have already seen the post do not go back and read the note. → Fact-checkers publicise their work, communicate retractions and even directly approach public figures to ask for corrections.
Psychological Harm	Argued → High cognitive load in viewing sensitive content, but protective policies are in place for welfare of experts. → Fact-checkers are professionally trained and accustomed to encountering sensitive content.	Not Discussed → No protective policies proposed for the cognitive load experienced in viewing sensitive content.	→ Psychological risks associated with laypeople encountering sensitive content are left undiscussed by social media platforms. → Lack of protective policies exposes platforms to legal liabilities. → That said, rowd-sourced moderation is voluntary, and users are not forced to consume/verify sensitive content.
Unpaid Labour	No → Third-party fact-checkers are hired by platforms to advise on handling of misleading content.	Yes → Relies on unpaid volunteers to engage in the task without labor protection.	→ Addressing misleading content is a demanding task, which deserves adequate remuneration. → Shifting responsibility to unpaid volunteers diminishes the fact-checkers' work and undermines the public's right to "the whole truth".

Table 1. Comparison of capabilities of *Third-party Fact-checker based Moderation* and *Community Notes* on social media platforms. Our analysis uses a three-point scale: Low capability in red, Mixed or Unclear capability in yellow, and High capability in green. We also outline the proposed advantages and drawbacks of both moderation approaches as argued by key stakeholders (e.g., social media platforms and fact-checkers), along with brief details on our critical evaluation of these claims.

verdict. We note a core epistemological difference in how the two approaches view facts: community notes as subjective constructs that are a matter of personal opinion, compared to fact-checking as an objective truth to be discovered through analyzing evidence[69,70]. This raises a fundamental question for platforms, policymakers, and the public: can democratic legitimacy and epistemic truth coexist in content moderation?

- **Bias:** Proponents of community notes have argued that the cross-perspective consensus required for publication means that these crowd-sourced fact-checks are inherently less politically biased than those written by professional fact-checkers. However, the majority of sources that are cited in notes are left-leaning news outlets rated to be of high factuality[71], suggesting that similar perspectives emerge in fact-checking and community notes. Moreover, notes for posts from US Republicans are proposed more often and rated as more helpful than for posts from US Democrats[72]. Community note users may also be susceptible to various cognitive biases: crowdworkers generally overestimate the truthfulness of claims, are overconfident in their abilities to judge the truthfulness of statements, and their truthfulness judgments are adversely impacted by their opinion of the claimant[73]. It has also been highlighted that while professional fact-checks are expected to adhere to formal and neutral communication styles, there is no such onus on community volunteers. Note writers may attempt to manipulate readers using highly persuasive but logically incoherent argumentation[74].
- **Transparency:** Platforms like X and Meta have open-sourced their community notes algorithms[28,75], enabling external inspection and reuse. In principle, this promotes transparency, letting users and readers understand how notes are assigned. While professional fact-checkers present evidence logically, their decision-making often involves intuition[14,63], as likely with community reviewers. Still, algorithmic transparency may boost trust. In practice, however, documentation lacks detail. For instance, ideological diversity is modeled along a single axis—adequate for political bias[17], but insufficient for capturing cultural, linguistic, or conspiratorial biases, requiring further study. Additionally, note authors remain anonymous, obscuring potential biases that could be inferred from past behavior. In contrast, fact-checking organizations follow strict transparency standards[76,77], often naming individual fact-checkers and disclosing sources, funding, and correction policies. Community notes lack a formal correction system; updates depend on users to detect and vote on inaccuracies.
- **Coordinated adversarial attacks:** Online social networks have long been susceptible to coordinated adversarial attacks — be it from automated bots or groups of users who inorganically inflate social reputations, amplify specific narratives and engage in other potentially harmful behaviors[78,79]. Recent studies emphasize that community moderation is also exposed to the same vulnerability[31,44]. Although platforms implement safeguards to validate contributors as real people and not adversarial actors, these mechanisms are not perfect and can be gamed with the right resources. Thus, entrusting a virtually unrestricted user base with content moderation responsibilities opens the door for malicious groups to exploit the system. By creating an artificial perception of internal disagreement through collusive inorganic activity on benign content, such groups can deceive the community notes algorithm and gain the platform's trust. This could place them in a dangerously advantageous position, allowing them to launch coordinated attacks on authentic narratives — suppressing their reach by casting doubt on their credibility through misleading community notes that falsely debunk the original content.

Do community notes help in countering misinformation?

There is insufficient data on the effectiveness of community notes as interventions for misinformation.

- **Impacts on misinformation beliefs:** The additional context provided by community notes appears to be appreciated by readers: they are judged as more trustworthy than simple labels that flag misinformation with no additional detail[32]. However, the study did not compare community notes with context written by professional fact-checkers, so it remains unclear whether a clear preference exists between the two approaches. Community notes and news article suggestions were found to be equally effective in reducing people's belief in and intention to repost misleading posts on social media[74]. An analysis of a single health-related claim in the same study found that community notes were more effective interventions for a positive framing (e.g., "This food may cure cancer"), whereas related articles were more effective for a negative framing (e.g., "This food may cause cancer"). It is unclear how far this may generalize to other claims and domains. Notes may also have negative tradeoffs: displaying community notes leads users to post more negative, angry, disgusted replies to misleading posts[80].
- **Impacts on misinformation spread:** Although Meta and X do not automatically reduce the reach or visibility of a post that has received a note[18], there is some evidence suggesting that notes are useful in curbing misinformation spread: posts on Twitter/X that are labeled by community notes as 'misleading' receive 37% fewer retweets than posts deemed 'not misleading'.[81] Community notes also increase the probability of tweet retractions and deletions and expedite the retraction of misleading posts.[33,82] On the other hand, community notes can draw attention to misleading posts, leading to increases in likes, engagement, and followers for accounts that receive 'misleading' community note labels.[62]

Are community notes an ethical replacement for professional fact-checkers?

Several ethical concerns have been raised with regard to an increasing reliance by social media platforms on crowd-sourced fact-checking.

- **Psychological harm:** Professional fact-checkers are accustomed to encountering harmful content (e.g., violence, abuse, and other explicit content) regularly, and receive dedicated training and support to manage the psychological consequences. Shifting this work to volunteers without any form of psychological support risks severe harm to those who write and rate community notes[83]. There is also a risk for platforms of potential liability and legal action by content moderators[84].
- **Unpaid labor:** In addition to the emotional burden borne by community moderators, relying solely on community notes raises the ethical issue of expecting non-professional volunteers to carry out intricate research on behalf of for-profit companies (distinct from contributors to non-profit Wikipedia) without remuneration or labor protection. The task of fact-checking requires skilled, challenging work, and those who undertake it should be fairly compensated for it.
- **De-funding and de-professionalization of fact-checking:** In addition to eroding fact-checking organizations' capacity to perform their basic functions, divestment in fact-checking devalues and diminishes the critical importance of access to reliable information for functioning societies[85]. Although community moderation relies on unpaid labor, it lacks professional standards, particularly in terms of training, accountability, and methodology. Community notes should be viewed as an approach that complements professional fact-checking rather than replacing it.

Recommendations

Drawing from our discussions thus far, it is clear that community moderation presents several unresolved issues that keep it from realizing its potential. Social media platforms have not yet addressed these issues when endorsing the community-driven approach as a replacement for expert-led moderation. We argue that platforms must improve the design of community moderation algorithms, taking into account their assumptions regarding the efficient user collaboration that is necessary for the success of this moderation approach. Here we put forth some recommendations for how the current issues with community moderation can be addressed by platforms and policymakers. Table 2 provides a summary of our discussion in this section.

Collaboration between community and experts

Community moderation represents an important step towards democratizing and scaling up content moderation. Yet we believe that its adoption as a replacement for fact-checkers is a missed opportunity for fruitful collaboration between experts and the crowd. Several reputed experts have advocated for such a collaborative approach[19, 84, 86]. Involving fact-checkers at various steps in the community moderation framework can overcome many of its unresolved issues:

- **Distributing the workload by filtering claims based on risk and ease of verification:** The crowd can address "low-hanging fruit" of mis- and disinformation. Previously, third-party fact-checking programs for content moderation have used machine learning solutions to address low-risk content, but with minimal human oversight. Community-driven content moderation allows volunteers to propose notes on relevant content by leveraging previously fact-checked claims and AI to surface already verified information—a technique that has proven effective in the mitigation of misleading content[82]. This, in turn, would allow professional fact-checkers to focus on emergent, high-risk claims that demand deeper investigation with an expert touch.
- **Fact-checkers acting as secondary reviewers to approve valid notes stuck in peer-review:** A collaborative approach can help alleviate bias-related issues and resolve cases where helpful notes do not reach consensus. Fact-checking experts can provide an expert, third-party assessment of the notes and content, analyzing both sides of the story where disputing opinions among laypersons might stall the peer review process. Earlier, some platforms had in-house teams of moderators who would take action against potentially harmful content, based on the advise of third-party fact-checkers. Now with the community notes model, platforms can, similarly, appoint further layer of governance comprising of established fact-checking experts. Through an internal content moderation management system, platforms can empower these experts to act as reviewers who can grant approval to helpful notes proposed by volunteers. This could be particularly useful for valid notes that exhibit some agreement within the community, but fall short of the acceptability threshold due to partisan objections. We note that research has experimented with using AI to help with this task.[87]
- **Flagging investigation worthy claims for professional fact-checkers:** The crowd could flag content in need of deeper analysis by expert fact-checkers. Fact-checkers often spend significant time just looking for claims that present the greatest potential risk for society. Having the crowd's support would help professional fact-checkers filter the vast stream of user-generated content to a smaller set of relatively high-risk claims. Platforms have always presented user-reported content to hired independent fact-checkers for verification. Community moderation presents a more transparent and democratic way of implementing such a flagging or priority system, though the concerns raised about transparency above

Category	Recommendation	Challenges Addressed	Details
Collaboration between community and experts	Workload distribution	• Breadth • Volume • Expertise	• Crowdwork to verify repetitive misinformation or widely debunked claims with ref • Fact-checkers to concentrate on new high-risk claims that need creation of new knowledge
	Fact-checkers as secondary reviewers	• Bias • Speed	• Fact-checkers can assess notes subjectively, considering all sides of the narrative • Fact-checkers to act as reviewers and approve notes that show partial helpfulness but do not meet the platform's threshold
	Flagging investigation-worthy claims	• Volume • Expertise • Bias	• Use community notes to identify and/or prioritize check-worthy claims for fact-checkers • Transparency through providing overview of distinct ideological groups writing and flagging posts
Collaboration between technology and the community	Social opinion analytics	• Speed • Adversarial attacks	• Expedite bridging diverse perspectives by identifying users more likely to engage in constructive debate • Improve robustness to brigading and/or coordinated attacks
	Fusing Community Notes	• Bias	• Identify points of agreement and discord between users • Generate notes that diverse perspectives can agree on
	AI-agents to simulate crowds	• Bias • Volume	• Flag posts with potentially sensitive content for review by experts
	Previously "community-noted posts"	• Volume	• Recommend notes from related posts • Enable cross-platform community moderation

Table 2. A summary of our recommendations to address the challenges faced by community notes. We present the list of our recommendations, the specific challenges they address, and a brief description of how they could be implemented within the community moderation algorithm.

still apply. This system could further be augmented with a transparent view of the users engaging with investigation-worthy content. Such contextual analysis, commonly included in expert fact-checks, helps illuminate patterns in how misinformation spreads and which groups are most affected.

- **Cross-platform community notes:** Part of the success in removing child sexual abuse material (CSAM) and other clearly illegal material is the existence of centralized, third-party repositories (e.g., The Internet Watch Foundation and GIFCT). No such resource linking fact-checks or community notes to content currently exists. One of the major strengths of X's Community Notes approach is its open source algorithm, which has been adopted by Meta[28]. A centralized community moderation system could be facilitated by multiple social media platforms, so notes that appear on one platform appear for the same claim on others. Such an approach could also help platforms with fewer users respond faster to misleading or other harmful content.

Collaboration between technology and the community

Alongside the collaboration between fact-checking experts and the crowd, we propose how technology could address some of the pitfalls of community moderation. AI and network analysis can improve the efficiency of key stages of the current community moderation, and even automate certain processes—ultimately enhancing the productivity of the crowd.

- **Auditing 'diverse' perspectives:** As we have discussed, the lack of transparency with regard to the users of community notes remains a challenge with potential for technical solutions. For example, proposing quantitative metrics to measure the diversity of perspectives, or developing AI techniques for improving the diversity of and de-biasing such clusters.

- **Using network analytics to address the partisan bottlenecks:** A major source of tension and debate among experts is how the algorithm handles the bridging of diverse perspectives in the peer review process. Reaching a consensus in community moderation on certain issues is not as easy as the platforms make it sound, evident from how valid notes are often left unpublished due to insufficient votes[47,65]. Research on modeling social opinion dynamics offers network analytics methods to efficiently identify users with opposing perspectives by studying their online activity[42,68]. Some studies even analyze how users with competing views could be selected from the network such that they would engage in a constructive debate on a particular topic[88,89]. Such methods could help address the major bottleneck of community moderation by finding people who are willing to rate notes that potentially oppose their own views.
- **Using network analytics to identify collusive groups:** The community notes algorithm is at risk from collusive behavior by malicious user groups. In particular, these groups may: *(i)* deceive the algorithm through manufactured internal disagreement to launch attacks on authentic content at opportune moments; or *(ii)* suppress helpful notes by rating them negatively simply because they conflict with the group's preferred narrative. Prior research on social network analysis has proposed frameworks for detecting collusive behavior on social media – such as groups of users artificially inflating social reputations or amplifying specific narratives[90]. Although early detection of artificial disagreement among community notes contributors is a fundamentally different task compared to existing efforts, we believe these frameworks could still offer useful foundations for spotting coordinated manipulation in community moderation. Further research in this sphere is critical. Researchers and platforms must continue analyzing community notes data to uncover collusive activities and the underlying patterns of user interaction to build robust defenses against such manipulation in community moderation.
- **Using AI to fuse information from various proposed community notes:** As discussed earlier, a significant number of valid notes remain unpublished because users with opposing viewpoints often fail to reach consensus on them[47]. Some studies suggest that this could partly be because the proposed notes present a biased narrative – one that aligns with the contributor's perspective, sometimes omitting key information. Moreover, this also causes "the whole truth" to be fragmented across multiple notes, with no single note providing a clarity on the situation. Consequently, users from diverse perspectives find it hard to rate a particular note as helpful. To address this, *Supernotes* have been proposed, which are AI-generated notes that integrate information from all proposed notes for a post into a single, cohesive version[87]. These *Supernotes* are designed to maximize the probability of consensus among users by reflecting historical patterns of writing in community notes rated as helpful by diverse user groups[87]. We believe that *Supernotes* and other similar methods hold great promise for community moderation by synthesizing notes that offer a comprehensive and neutral account of events – making them more acceptable across a broad range of user perspectives.
- **AI agent-augmented crowds for proposing community notes:** With recent advancements in reasoning capabilities of LLMs, agentic fact-checking has shown great promise for verifying real-world-claims. Parallelly, as community notes have become prominent, some studies have also advocated for the potential of LLM-agents to simulate crowds for community moderation[91]. These swarms of agents have demonstrated an ability to classify the truthfulness of social media posts, with performance comparable to that of human crowds. However, it is important to keep in mind that these studies were performed in controlled environments, but real-world, user-generated data is much more noisy and complex. Further research is needed to evaluate the performance of such systems in practical, real-time scenarios. That said and building on insights from prior efforts in developing successful agentic fact-checking systems, we believe that LLM agents could be useful for proposing notes on problematic content. And as community notes grow, these agents could be further tuned to align with the performance of ideal human crowd-workers, particularly to overcome common limitations of human crowds—such as truthfulness overestimation and cognitive biases[73]. Note that, we do not advocate for fully replacing human community notes contributors. Rather, we envision these AI-agents to augment human efforts in proposing community notes by automating certain parts of the process.
- **Using AI to suggest similar previously 'community-noted' posts:** A key step in automatic fact-checking is to identify whether a seemingly new claim can be verified using a previously fact-checked claim[60,92]. A similar strategy could be adopted for community notes. Such a system could also be augmented to enable *cross-platform community moderation*. Given the transparent and open-source nature of community notes, platforms adopting it could collaborate to centralize these notes. AI could then match new posts to previously "community-noted" posts, minimizing redundant verification efforts from the community across platforms.

Conclusion

Community notes offer significant promise for addressing misinformation on social media platforms, with the potential to increase the speed, scale, and participation in fact-checking. Yet our analysis of the current implementations of the community note model reveals critical limitations of crowdsourced fact-checking and demonstrates that they fall well short of a comprehensive solution to the socio-technical challenges of misinformation, and to the complex issues of bias and trust in information. In this paper, we have outlined how some of these challenges might be addressed. Collaborating with fact-checkers and journalists can complement a community moderation system and provide the necessary expertise and accuracy crucial in complex and high-stakes contexts. Technical approaches can support the transparency of this approach and bolster resilience to bias and malign manipulation. Our recommendations for the future of community notes point to how such a model can scale the reach and participation in fact-checking. Community-driven efforts are necessary, but not sufficient alone as a means of combating misinformation. Social media platforms must take responsibility for foregrounding and incentivizing high-quality contributions, while policymakers can mandate transparency and common standards among platforms, to ensure that systems are designed to serve the broader information society. Community moderation may help democratize fact checking, but without the integration of expert viewpoints, algorithmic transparency, and institutional support, it risks consolidating consensus over establishing correctness.

Author contributions statement

Conceptualized by I.A., T.C. and P.N.; initial manuscript prepared by D.S., G.W., I.A., T.C, P.N.; all authors contributed to finalizing the manuscript. The author names are arranged alphabetically by last name.

References

1. Mosseri, A. & Meta. Addressing hoaxes and fake news (2016). Accessed: 29 April 2025, https://about.fb.com/news/2016/12/news-feed-fyi-addressing-hoaxes-and-fake-news/.
2. TechCoalition. Working together to end online child sexual exploitation and abuse. Accessed: 29 April 2025, https://www.technologycoalition.org.
3. TikTok. Guidelines: Integrity and authenticity. Accessed: 29 April 2025, https://www.tiktok.com/community-guidelines/en/integrity-authenticity.
4. YouTube. Community guidelines. Accessed: 29 April 2025, https://www.youtube.com/howyoutubeworks/policies/community-guidelines/#enforcing-community-guidelines.
5. TikTok. How enforcement technology works (2024). Accessed: 29 April 2025, https://transparency.meta.com/en-gb/enforcement/detecting-violations/how-enforcement-technology-works/.
6. TikTok. Community guidelines enforcement report (2025). Accessed: 29 April 2025, https://www.tiktok.com/transparency/en/community-guidelines-enforcement-2024-4.
7. Gorwa, R., Binns, R. & Katzenbach, C. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Soc.* **7**, 2053951719897945, DOI: 10.1177/2053951719897945 (2020). https://doi.org/10.1177/2053951719897945.
8. Horta Ribeiro, M., Cheng, J. & West, R. Automated content moderation increases adherence to community guidelines. In *Proceedings of the ACM Web Conference 2023*, WWW '23, 2666–2676, DOI: 10.1145/3543507.3583275 (Association for Computing Machinery, New York, NY, USA, 2023).
9. Tonneau, M. *et al.* Hateday: Insights from a global hate speech dataset representative of a day on twitter (2024). 2411.15462.
10. DeVerna, M. R., Yan, H. Y., Yang, K.-C. & Menczer, F. Fact-checking information from large language models can decrease headline discernment. *Proc. Natl. Acad. Sci.* **121**, e2322823121, DOI: 10.1073/pnas.2322823121 (2024). https://www.pnas.org/doi/pdf/10.1073/pnas.2322823121.
11. Zhou, X., Sharma, A., Zhang, A. X. & Althoff, T. Correcting misinformation on social media with a large language model (2024). 2403.11169.
12. Augenstein, I. *et al.* Factuality challenges in the era of large language models and opportunities for fact-checking. *Nat. Mach. Intell.* **6**, 852–863, DOI: 10.1038/s42256-024-00881-z (2024).
13. Micallef, N., Armacost, V., Memon, N. & Patil, S. True or false: Studying the work practices of professional fact-checkers. *Proc. ACM Hum.-Comput. Interact.* **6**, DOI: 10.1145/3512974 (2022).
14. Warren, G., Shklovski, I. & Augenstein, I. Show me the work: Fact-checkers' requirements for explainable automated fact-checking. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, 1–21, DOI: 10.1145/3706598.3713277 (ACM, 2025).
15. Nakov, P. *et al.* Automated fact-checking for assisting human fact-checkers. In Zhou, Z.-H. (ed.) *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, 4551–4558, DOI: 10.24963/ijcai.2021/619 (International Joint Conferences on Artificial Intelligence Organization, 2021). Survey Track.

16. Kolla, M., Salunkhe, S., Chandrasekharan, E. & Saha, K. LLM-Mod: Can large language models assist content moderation? In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, DOI: 10.1145/3613905.3650828 (Association for Computing Machinery, New York, NY, USA, 2024).

17. Wojcik, S. *et al.* Birdwatch: Crowd wisdom and bridging algorithms can inform understanding and reduce the spread of misinformation. *arXiv preprint arXiv:2210.15723* DOI: 10.48550/arXiv.2210.15723 (2022).

18. Kaplan, J. More speech and fewer mistakes (2025). Accessed: 29 April 2025, https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/.

19. The International Fact-Checking Network. An open letter to Mark Zuckerberg from the world's fact-checkers, nine years later (2025). Accessed: 29 April 2025, https://www.poynter.org/ifcn/2025/an-open-letter-to-mark-zuckerberg-from-the-worlds-fact-checkers-nine-years-later/.

20. Surowiecki, J. *The Wisdom of Crowds* (Anchor, 2005).

21. Wikipedia Volunteers. Help: Editing (2024). Accessed: 7 May 2025, https://en.wikipedia.org/wiki/Help:Editing.

22. Reddit. Moderator code of conduct (2024). Accessed: 7 May 2025, https://redditinc.com/policies/moderator-code-of-conduct.

23. Halprin, M. & O'Connor, J. F. On policy development at YouTube (2022). Accessed: 29 April 2025, https://blog.youtube/inside-youtube/policy-development-at-youtube/.

24. Coleman, K. Introducing Birdwatch, a community-based approach to misinformation (2021). Accessed: 29 April 2025, https://blog.x.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation.

25. X. About community notes on X. Accessed: 29 April 2025, https://help.x.com/en/using-x/community-notes.

26. Allen, J., Martel, C. & Rand, D. G. Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in twitter's birdwatch crowdsourced fact-checking program. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, DOI: 10.1145/3491102.3502040 (Association for Computing Machinery, New York, NY, USA, 2022).

27. Coleman, K. Introducing birdwatch, a community-based approach to misinformation. Accessed: 23 May 2025, https://blog.x.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation?utm_source=chatgpt.com.

28. Meta. Testing begins for community notes on Facebook, Instagram and Threads (2025). Accessed: 29 April 2025.

29. Weibo. Community "notes" function launch announcement (2023). Accessed: 29 April 2025, https://weibo.com/1934183965/NfN4bB4ZE/.

30. Presser, A. Testing a new feature to enhance content on tiktok (2025). Accessed: 29 April 2025, https://newsroom.tiktok.com/en-us/footnotes/.

31. Saeed, M., Traub, N., Nicolas, M., Demartini, G. & Papotti, P. Crowdsourced fact-checking at Twitter: How does the crowd compare with experts? In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, 1736–1746, DOI: 10.1145/3511808.3557279 (Association for Computing Machinery, New York, NY, USA, 2022).

32. Drolsbach, C. P., Solovev, K. & Pröllochs, N. Community notes increase trust in fact-checking on social media. *PNAS Nexus* DOI: 10.1093/pnasnexus/pgae217 (2024).

33. Renault, T., Amariles, D. R. & Troussel, A. Collaboratively adding context to social media posts reduces the sharing of false news, DOI: 10.48550/arXiv.2404.02803 (2024).

34. Mutsvairo, B. & Salgado, S. Is citizen journalism dead? an examination of recent developments in the field. *Journalism* **23**, 354–371, DOI: 10.1177/1464884920968440 (2022).

35. Splichal, S. & Dahlgren, P. Journalism between de-professionalisation and democratisation. *Eur. J. Commun.* **31**, 5–18, DOI: 10.1177/0267323115614196 (2016).

36. X. Dsa transparency report - october 2024. Accessed: 23 May 2025, https://transparency.x.com/dsa-transparency-report.html.

37. Meta. How meta enforces its policies. Accessed: 23 May 2025, https://transparency.meta.com/enforcement/.

38. X Corp. & Masjoody, M. X corp. v. masjoody, 2025 bcca 89. Accessed: 12 May 2025, https://www.canlii.org/en/bc/bcca/doc/2025/2025bcca89/2025bcca89.html.

39. Dave Patrick Underwood, M. Dave patrick underwood v. meta platforms, inc. Accessed: 12 May 2025, https://www.documentcloud.org/documents/21174499-underwood-v-meta-complaint/.

40. X Corp. Global transparency report h2 2024. Accessed: 12 May 2025, https://transparency.x.com/en/reports/global-reports/2025-transparency-report.

41. Wikipedia Volunteers. List of edit wars on wikipedia. Accessed: 14 May 2025, https://en.wikipedia.org/wiki/List_of_edit_wars_on_Wikipedia.

42. Sasahara, K. *et al.* Social influence and unfollowing accelerate the emergence of echo chambers. *J. Comput. Soc. Sci.* **4**, 381–402, DOI: 10.1007/s42001-020-00084-7 (2021).

43. Arora, S. D., Singh, G. P., Chakraborty, A. & Maity, M. Polarization and social media: A systematic review and research agenda. *Technol. Forecast. Soc. Chang.* **183**, 121942, DOI: https://doi.org/10.1016/j.techfore.2022.121942 (2022).

44. Martel, C., Allen, J., Pennycook, G. & Rand, D. G. Crowds can effectively identify misinformation at scale. *Perspectives on Psychol. Sci.* **19**, 477–488, DOI: 10.1177/17456916231190388 (2023). PMID: 37594056, https://doi.org/10.1177/17456916231190388.

45. Wikipedia. Community engagement insights/2018 report (2018). https://web.archive.org/web/20190724142049/https://meta.wikimedia.org/wiki/Community_Engagement_Insights/2018_Report.

46. Wack, M., Duskin, K. & Hodel, D. Political fact-checking efforts are constrained by deficiencies in coverage, speed, and reach (2024). 2412.13280.

47. Center for Countering Digital Hate. Rated Not Helpful: How X's Community Notes system falls short on election disinformation (2024). https://counterhate.com/research/rated-not-helpful-x-community-notes/.

48. Bhat, P. & Klein, O. Covert hate speech: White nationalists and dog whistle communication on twitter. *Twitter, public sphere, chaos online deliberation* 151–172, DOI: 10.1007/978-3-030-41421-4_7 (2020).

49. Martel, C. & Rand, D. G. Fact-checker warning labels are effective even for those who distrust fact-checkers. *Nat. Hum. Behav.* **8**, 1957–1967, DOI: 10.1038/s41562-024-01973-x (2024).

50. Martel, C. & Rand, D. G. Misinformation warning labels are widely effective: A review of warning effects and their moderating features. *Curr. Opin. Psychol.* **54**, 101710, DOI: https://doi.org/10.1016/j.copsyc.2023.101710 (2023).

51. Borenstein, N., Warren, G., Elliott, D. & Augenstein, I. Can community notes replace professional fact-checkers? (2025). 2502.14132.

52. X. Locking and unlocking the ability to write notes. Accessed: 29 April 2025, https://communitynotes.x.com/guide/en/contributing/writing-ability.

53. Stencel, M., Ryan, E. & Luther, J. With half the planet going to the polls in 2024, fact-checking sputters (2024). https://reporterslab.org/with-half-the-planet-going-to-the-polls-in-2024-fact-checking-sputters/.

54. Balod, H. S. S. & Hameleers, M. Fighting for truth? the role perceptions of filipino journalists in an era of mis-and disinformation. *Journalism* **22**, 2368–2385 (2021).

55. Cheruiyot, D. & Ferrer-Conill, R. "fact-checking africa" epistemologies, data and the expansion of journalistic discourse. *Digit. Journalism* **6**, 964–975 (2018).

56. Ababakirov, A. *et al. Meeting the challenges of information disorder in the Global South* (2022).

57. Vinhas, O. & Bastos, M. The weird governance of fact-checking and the politics of content moderation. *new media & society* 14614448231213942 (2023).

58. The International Fact-Checking Network. Global Fact Check Fund awards $2 million to 20 fact-checking groups across 15 countries (2025). Accessed: 19 May 2025, https://www.poynter.org/ifcn/2025/global-fact-check-fund-awards-2-million-to-20-fact-checking-groups-across-15-countries/.

59. Neumann, T., De-Arteaga, M. & Fazelpour, S. Justice in misinformation detection systems: An analysis of algorithms, stakeholders, and potential harms. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 1504–1515, DOI: 10.1145/3531146.3533205 (Association for Computing Machinery, New York, NY, USA, 2022).

60. Nakov, P. *et al.* The CLEF-2021 CheckThat! Lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43*, 639–649 (Springer, 2021).

61. Graham, M., Straumann, R. K. & Hogan, B. Digital divisions of labor and informational magnetism: Mapping participation in wikipedia. *Annals Assoc. Am. Geogr.* **105**, 1158–1178 (2015).

62. Wirtschafter, V. & Majumder, S. Future challenges for online, crowdsourced content moderation: Evidence from Twitter's community notes. *J. Online Trust. Saf.* **2**, DOI: 10.54501/jots.v2i1.139 (2023).

63. Graves, L. Anatomy of a fact check: Objective practice and the contested epistemology of fact checking. *Commun. culture & critique* **10**, 518–537, DOI: https://doi.org/10.1111/cccr.12163 (2017).

64. Zhao, A. & Naaman, M. Insights from a comparative study on the variety, velocity, veracity, and viability of crowdsourced and professional fact-checking services. *J. Online Trust. Saf.* **2**, DOI: 10.54501/jots.v2i1.118 (2023).

65. Mahadevan, A. Meta will attempt crowdsourced fact-checking. here's why it won't work (2025). Accessed: 29 April 2025, https://www.poynter.org/commentary/2025/meta-community-notes-crowdsourced-fact-checking-x/.

66. Pröllochs, N. Community-Based Fact-Checking on Twitter's Birdwatch Platform. *Proc. Int. AAAI Conf. on Web Soc. Media* **16**, 794–805, DOI: 10.1609/icwsm.v16i1.19335 (2022).

67. Solovev, K. & Pröllochs, N. References to unbiased sources increase the helpfulness of community fact-checks (2025). 2503.10560.

68. Yasseri, T. & Menczer, F. Can crowdsourcing rescue the social marketplace of ideas? *Commun. ACM* **66**, 42–45, DOI: 10.1145/3578645 (2023).

69. Uscinski, J. E. & Butler, R. W. The epistemology of fact checking. *Critical Rev.* **25**, 162–180 (2013).

70. Amazeen, M. A. Revisiting the epistemology of fact-checking. *Critical review* **27**, 1–22 (2015).

71. Kangur, U., Chakraborty, R. & Sharma, R. Who checks the checkers? exploring source credibility in Twitter's community notes (2024). 2406.12444.

72. Renault, T., Mosleh, M. & Rand, D. G. Republicans are flagged more often than democrats for sharing misinformation on x's community notes, DOI: 10.31234/osf.io/vk5yj_v4 (2025).

73. Draws, T. *et al.* The effects of crowd worker biases in fact-checking tasks. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 2114–2124, DOI: 10.1145/3531146.3534629 (Association for Computing Machinery, New York, NY, USA, 2022).

74. Kankham, S. & Hou, J.-R. Community notes vs. related articles: Assessing real-world integrated counter-rumor features in response to different rumor types on social media. *Int. J. Human-Computer Interact.* 1–15, DOI: 10.1080/10447318.2024.2400389 (2024).

75. X community notes. https://github.com/twitter/communitynotes. Accessed: 2025-04-20.

76. International Fact-Checking Network. The commitments of the code of principles (2016). https://ifcncodeofprinciples.poynter.org/the-commitments.

77. European Fact-Checking Standards Network. Code of standards (2022). https://efcsn.com/code-of-standards/.

78. Pacheco, D. *et al.* Uncovering coordinated networks on social media: Methods and case studies. In *Proc. International AAAI Conference on Web and Social Media (ICWSM)*, vol. 15, 455–466, DOI: 10.1609/icwsm.v15i1.18075 (2021).

79. Bradshaw, S. & Howard, P. Troops, trolls and troublemakers: A global inventory of organized social media manipulation. *Comput. propaganda research project* (2017).

80. Chuai, Y., Sergeeva, A., Lenzini, G. & Pröllochs, N. Community fact-checks trigger moral outrage in replies to misleading posts on social media, DOI: 10.48550/arXiv.2409.08829 (2024). ArXiv:2409.08829 [cs].

81. Drolsbach, C. P. & Pröllochs, N. Diffusion of Community Fact-Checked Misinformation on Twitter. *Proc. ACM on Human-Computer Interact.* **7**, 1–22, DOI: 10.1145/3610058 (2023).

82. Gao, Y., Zhang, M. & Rui, H. Can Crowdchecking Curb Misinformation? Evidence from Community Notes, DOI: 10.2139/ssrn.4992470 (2024).

83. Steiger, M., Bharucha, T. J., Venkatagiri, S., Riedl, M. J. & Lease, M. The psychological well-being of content moderators: the emotional labor of commercial moderation and avenues for improving support. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, 1–14 (2021).

84. Gibert, S. Three reasons Meta will struggle with community fact-checking (2025). Accessed: 12th May 2025, https://www.technologyreview.com/2025/01/29/1110630/three-reasons-meta-will-struggle-with-community-fact-checking.

85. Moran, R. E., Schafer, J., Bayar, M. & Starbird, K. The end of trust and safety?: Examining the future of content moderation and upheavals in professional online safety efforts. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, DOI: 10.1145/3706598.3713662 (Association for Computing Machinery, New York, NY, USA, 2025).

86. Holan, A. D. Will the future of fact-checking flourish or founder? 2025 marks a new turning point (2025). Accessed: 29 April 2025, https://www.poynter.org/fact-checking/2025/angie-drobnic-holan-international-fact-checking-day.

87. De, S., Baxter, J., Bakker, M. & Saveski, M. Supernotes: Driving consensus in crowd-sourced fact-checking. In *Proceedings of the ACM Web Conference 2025 (WWW '25)*, 11, DOI: 10.1145/3696410.3714934 (ACM, New York, NY, USA, 2025).

88. Garimella, K., De Francisci Morales, G., Gionis, A. & Mathioudakis, M. Reducing controversy by connecting opposing views. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, 81–90, DOI: 10.1145/3018661.3018703 (Association for Computing Machinery, New York, NY, USA, 2017).

89. Garimella, K., De Francisci Morales, G., Gionis, A. & Mathioudakis, M. Factors in recommending contrarian content on social media. In *Proceedings of the 2017 ACM on Web Science Conference*, WebSci '17, 263–266, DOI: 10.1145/3091478.3091515 (Association for Computing Machinery, New York, NY, USA, 2017).

90. Dutta, H. S. & Chakraborty, T. Blackmarket-driven collusion on online media: A survey (2020). 2008.13102.

91. Costabile, L., Orlando, G. M., Gatta, V. L. & Moscato, V. Assessing the potential of generative agents in crowdsourced fact-checking (2025). 2504.19940.

92. Shaar, S., Babulkov, N., Da San Martino, G. & Nakov, P. That is a known lie: Detecting previously fact-checked claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3607–3618, DOI: 10.18653/v1/2020.acl-main.332 (Association for Computational Linguistics, Online, 2020).