



Research Article

Professional and community-based fact-checking show different strengths, but neither performs strongly across trust, scalability, and impact

Fact-checking on social media matters more than ever. When Meta abandoned professional fact checkers in 2025 in favour of crowd-sourced annotations, it bet that community participation could substitute for institutional expertise. Our systematic review of 21 empirical studies of social media fact-checking (2022–2025) suggests the picture is more complicated than it appears. Community-based fact-checking is faster and scalable, whereas professional fact checkers are generally more trusted but often slow. Neither model, on its own, performs strongly across trust, scalability, and impact. The evidence suggests that hybrid systems combining community speed with expert verification are likely to be the most effective approach.

Authors: Petter Bae Brandtzaeg (1), Silje Susanne Alvestad (2,3), Jonas R. Kunst (4,5), Asbjørn Følstad (6)

Affiliations: (1) Department of Media and Communication, University of Oslo, Norway, (2) Language and Intelligence Academy, Norwegian Defence University College, Norway, (3) Department of Literature, Area Studies and European Languages, University of Oslo, Norway, (4) Department of Communication and Culture, BI Norwegian Business School, Norway, (5) Department of Psychology, University of Oslo, Norway, (6) SINTEF Digital, Norway

How to cite: Brandtzaeg, P. B., Alvestad, S. S., Kunst, J. R., & Følstad, A. (2026). Professional and community-based fact-checking show different strengths, but neither performs strongly across trust, scalability, and impact. *Harvard Kennedy School (HKS) Misinformation Review*, 7(4).

Received: April 8th, 2026. Accepted: July 22nd, 2026. Published: July 30th, 2026.

Research questions

- How do professional and community-based fact-checking systems on social media compare in terms of user trust?
- How do they compare in terms of scalability, including both processing volume and audience reach on social media?
- How do they compare in terms of impact on misinformation beliefs and sharing behavior on social media?
- What do potential differences imply for the design of more effective social media fact-checking interventions?

Essay summary

- We conducted a PRISMA-guided systematic review of 21 peer-reviewed empirical studies

¹ A publication of the Shorenstein Center on Media, Politics and Public Policy at Harvard University, John F. Kennedy School of Government.

published between 2022 and 2025 (i.e., the period capturing the emergence of community-based fact-checking systems in mainstream platform governance), identified through searches in ACM Digital Library, Web of Science, Scopus, and JSTOR.

- The studies were compared using a three-part framework focused on trust, scalability, and impact.
- Neither professional nor community-based fact-checking on social media performs strongly across all three dimensions at the same time.
- Professional fact-checking tends to perform better on trust and shows moderate impact but remains limited by low to moderate scalability.
- Community-based fact-checking seems to perform better on scalability, but its trust and impact remain moderate because visibility, partisanship, and selective uptake limit its corrective effects.
- The key problem is not that one model works and the other fails, but that current systems optimize different dimensions of fact-checking effectiveness on social media.
- Hybrid systems combining crowd-based speed with expert validation may offer a more promising path than treating professional and community-based fact-checking as direct substitutes.

Implications

Fact-checking on social media plays a crucial role in identifying and correcting misinformation (Lee et al., 2023). At the same time, social media has changed how people evaluate information and determine what to trust. The ability to search, compare sources, and investigate claims independently has made professional expertise one source of authority among many (Bartsch et al., 2025). One manifestation of this broader shift was Meta's 2025 decision to replace third-party professional fact checkers with crowd-sourced annotations. These developments have intensified debates over the relative strengths and limitations of different fact-checking models. This systematic review compares professional and community-based fact-checking across three dimensions central for fact-checking effectiveness on social media: trust, scalability, and impact. Although professional and community-based fact-checking should not be treated as mutually exclusive categories, the findings show that these models involve distinct trade-offs. Professional fact-checking seems to generate higher levels of trust and offers clearer verification standards, but it is also slower and less scalable. Community-based fact-checking can operate with greater speed and scale, but its corrective effects are more uneven because they depend on participation, visibility, user acceptance, and the risk of ideological bias in who contributes and which claims are corrected. Based on these findings, we identify three main implications.

Implications for platform design

Social media platforms often present community-based fact-checking as a scalable solution to the limits of professional fact-checking (Meta, 2025). Our review suggests that this claim should be treated cautiously. Community-based fact-checking clearly improves processing volume and can annotate misleading content much faster than professional fact-checking (Allen et al., 2021; Saeed et al., 2022; Zhao & Naaman, 2023). However, scalability at the point of annotation is not the same as scalability at the point of correction. Many community notes are never shown publicly because they fail to receive sufficient support from contributors representing diverse perspectives. (La Barbera et al., 2024; Pröllochs, 2022). Even visible notes may appear too late to interrupt early spread (Chuai et al., 2024), and they do not persuade all users equally (Allen et al., 2022; Kuuse et al., 2025). Community-based fact-checking should therefore not be treated as a self-sufficient replacement for professional expertise. Platforms should invest more in how corrections are surfaced and report what proportion of submitted notes become

visible and are actually seen by users. This matters especially in engagement-driven environments where misleading and AI-generated content can spread rapidly (Cotter, 2021; Metzler & Garcia, 2023; Schroeder et al., 2026).

Implications for policymakers and regulators

Public debate often frames fact-checking as either a censorship issue or a free-speech issue (Lewandowsky, 2025). Our findings suggest that this framing is too narrow. Fact-checking is also part of the epistemic infrastructure of digital public life. Across the reviewed studies, the central issue is whether corrective systems generate trust, achieve sufficient visibility, and produce meaningful impact. In this sense, a system that handles many posts but reaches only a small share of exposed users may satisfy formal expectations of participation while still having limited corrective value (Liu et al., 2025; Pilarski et al., 2024; Roozenbeek et al., 2024; Seelam et al., 2024). From this perspective, it is not enough to ask whether a system is open or participatory. It is also necessary to ask whether it works. Regulators should, therefore, look beyond process and consider functional effectiveness.

Implications for fact-checking organisations

Platform shifts toward community-based fact-checking should not be interpreted as evidence that professional fact-checking has become irrelevant. The reviewed literature suggests the opposite. Professional fact checkers retain a comparative advantage in trust and in the procedural rigor and domain-specific expertise that help sustain it (Liu et al., 2025; Primig, 2022; Zhao & Naaman, 2023). This remains especially important in high-stakes domains such as health and public affairs, where the costs of weak or inaccurate correction may be substantial (Allen et al., 2024; Zhao & Naaman, 2023). At the same time, professional fact-checking is unlikely to compete on speed and volume alone. Its most viable role may therefore lie in hybrid systems, where community participants help identify and annotate content at scale while professional fact checkers provide expert review of high-reach or high-stakes claims (Pilarski et al., 2024; Saeed et al., 2022; Zhao & Naaman, 2023).

Finally, impact should not be equated with volume. A correction system is only effective if people see it, trust it, and respond to it (Roozenbeek et al., 2024). Trust, scalability, and impact are linked, but they are not interchangeable. A system may scale without producing meaningful correction, while a trusted system may still fail to reach enough people to matter at the population level (Kyriakidou et al., 2022; Liu et al., 2025). Fact-checking also has to work inside engagement-driven platforms. Misleading content can generate attention, interaction, and advertising value, so platforms may face a tension between reducing misinformation and maximizing engagement (Metzler & Garcia, 2023; Schroeder et al., 2026). For this reason, platforms should report not only how many corrections are produced, but also how often users actually see them and whether the corrections reduce exposure to and sharing of misinformation. Platforms may reduce accusation of ideological bias by using transparent labeling criteria, consistent application across political viewpoints, and independent oversight where possible. Such functions could make fact-checking more accountable and less easy to dismiss as arbitrary or partisan. This matters even more as AI-generated content is likely to increase both the volume and speed of misleading content (Bashardoust et al., 2024; Schroeder et al., 2026). The practical challenge, then, is not to choose between professional and community-based fact-checking but to develop hybrid systems better aligned with trust, scalability, and impact within the same corrective model.

Findings

Finding 1: No single model performs strongly across trust, scalability, and impact at the same time.

Across the 21 studies reviewed, the clearest pattern is that trust, scalability, and impact do not align strongly within any current fact-checking model (see Appendix A for details). Professional fact-checking tends to show moderate trust and moderate impact but low to moderate scalability. Community-based fact-checking tends to allow high scalability but only moderate trust and moderate impact (Ejaz et al., 2025; Liu et al., 2025; Primig, 2022; Saeed et al., 2022; Zhao & Naaman, 2023). The problem is therefore not simply that one model works and the other does not. Rather, each model is built around a different design logic. Third-party professional fact-checking prioritizes procedural rigor and credibility, while community-based approaches prioritize speed and distributed participation. Neither model performs strongly across all three dimensions at once (Drolsbach et al., 2024; He et al., 2025; Zhou et al., 2025), suggesting that current fact-checking systems are structured around trade-offs rather than comprehensive effectiveness. Importantly, we view these models as analytical ideal types rather than mutually exclusive categories. Several reviewed studies suggest that hybrid arrangements may be more viable, combining community-based identification and annotation at scale with professional review of high-reach or high-stakes claims (Pilarski et al., 2024; Saeed et al., 2022; Zhao & Naaman, 2023).

Finding 2: Community-based fact-checking is faster and more scalable, but this does not reliably produce stronger correction.

The strongest comparative advantage of community-based fact-checking is scalability. Systems such as Community Notes can process large volumes of content and often do so much faster than professional fact checkers (Chuai et al., 2024; Saeed et al., 2022; Zhao & Naaman, 2023). However, scalability does not automatically translate into stronger impact. Many notes remain invisible because they fail to meet consensus thresholds (La Barbera et al., 2024; Pröllochs, 2022), others appear too late to alter early diffusion (Chuai et al., 2024), and coverage of topics can remain uneven (Pilarski et al., 2024). Community-based systems also appear vulnerable to ideological or partisan skew, both in participation and in corrective outcomes (Allen et al., 2022; Drolsbach et al., 2024; Kuuse et al., 2025; Pröllochs, 2022; Saeed et al., 2022). One comparative study also found no significant difference in misinformation reduction between professional fact-checking and Community Notes, despite the latter's much greater scalability (Liu et al., 2025). Taken together, this reveals that speed and reach alone are insufficient unless corrections also achieve visibility, timeliness, user acceptance, and sufficient protection against ideological or partisan distortion.

Finding 3: Trust is fragile in both models and depends on source identity, perceived neutrality, and framing.

Trust is central to the success of any fact-checking system, but it is fragile in both models. For professional fact checkers, trust often depends on institutional reputation, prior media trust, and perceptions of political neutrality (Park, 2023; Primig, 2022). Distrust is common where fact checkers are seen as elite, partisan, or aligned with mainstream institutions (Ejaz et al., 2025; Lavilles et al., 2023). Community-based fact-checking faces a different trust problem. Their participatory structure can make it appear more open and transparent, and some studies find that it is trusted more than simple misinformation flags (Drolsbach et al., 2024). But trust remains highly conditional when users suspect ideological skew or biased participation (Allen et al., 2022; Kuuse et al., 2025; Pretus et al., 2024). Across both models, framing also matters. Corrections framed as confirmation rather than direct refutation can generate greater

engagement and less resistance (Aruguete et al., 2024; Aruguete et al., 2025). This implies that trust is not built by sources alone but also by how corrections are rhetorically presented.

Finding 4: Both models show moderate impact, but the effects are often short-lived and unevenly distributed.

The overall evidence for impact is best described as moderate (see Appendix B). Professional fact-checking often reduces belief in false claims, especially among audiences predisposed to trust the source (Ejaz et al., 2025; Porter & Wood, 2021). Community-based fact-checking also shows measurable effects, including fewer reposts, greater deletion of misleading posts, and better recognition of misleading content (Chuai et al., 2024; Drolsbach & Pröllochs, 2022; Drolsbach et al., 2024). At the same time, many studies focus on short-term engagement or belief outcomes rather than longer-term behavioral change. Others suggest that fact-checking remains peripheral in everyday information routines (Kyriakidou et al., 2022). The main challenge is therefore not strong backfire effects, which appear to be rare, but the difficulty of getting corrections to the right users, at the right time, in a form they are willing to accept (Nyhan, 2021; Swire-Thompson et al., 2022). Overall, the findings point to a structural tension of current fact-checking systems: no existing model aligns trust, scalability, and impact in a consistently strong way. Professional and community-based approaches each address part of the problem, but neither fully resolves it. This suggests that the future of fact-checking may depend less on choosing between models than on finding ways to combine their respective strengths.

Methods

We conducted a PRISMA-guided systematic literature review of recent empirical studies (Page et al., 2020). The review focused on human-driven professional and community-based fact-checking. As detailed in Figure 1, we searched four academic databases, removed irrelevant records during title and abstract screening, and assessed the remaining full texts against predefined inclusion criteria. The four databases were selected to balance coverage across communication, psychology, journalism, HCI, information science, and computational social science: ACM Digital Library captures HCI and platform-governance research; Web of Science and Scopus provide broad interdisciplinary coverage; and JSTOR adds coverage of social science and media studies journals that may not be indexed as strongly in technical databases.

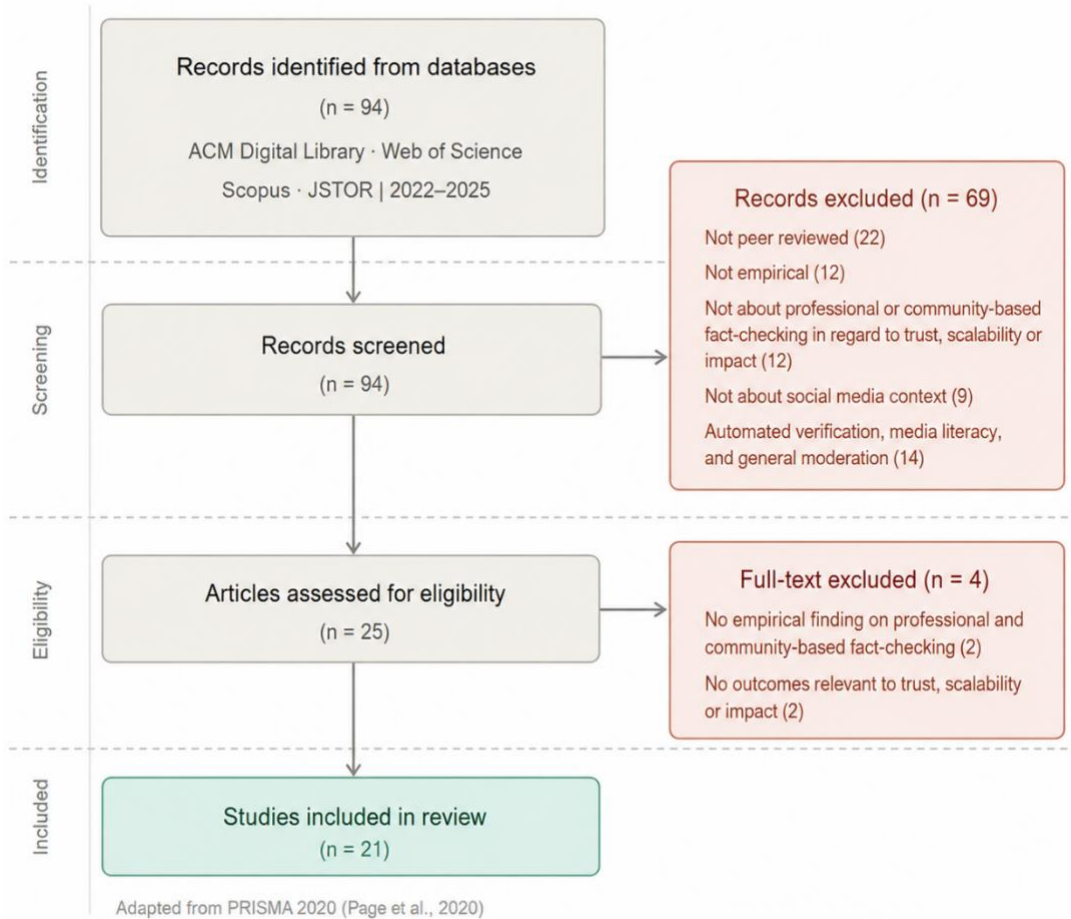


Figure 1. PRISMA flow diagram showing the identification, screening, exclusion, and inclusion of studies in the review.

Search strings were constructed by combining three groups of terms with Boolean operators (see also Appendix C). Terms within each group were combined with OR, while the three groups were combined with AND. The search string was: (“misinformation” OR “disinformation”) AND (“social media” OR “platform*” OR “Twitter” OR “X” OR “Facebook” OR “WhatsApp”) AND (“fact-checking” OR “fact checking” OR “fact-check*” OR “misinformation correction” OR “Community Notes” OR “Birdwatch” OR “community fact-checking” OR “professional fact-checking” OR “crowdsourced verification” OR “peer fact-checking” OR “expert verification”).

The search was limited to studies published between 2022 and 2025 because the empirical study of community-based fact-checking as an operational platform-governance system, especially after Twitter’s Birdwatch/Community Notes, became available for empirical analysis in this period (Pröllochs, 2022; Saeed et al., 2022). Earlier work on crowd-based accuracy judgments is relevant (e.g., Allen et al., 2021; Godel et al., 2021). However, these studies were not included in the final sample because they examined crowd-based assessment models rather than deployed social media fact-checking systems with public annotations, visible contributor participation, and platform-level correction processes such as Birdwatch/Community Notes.

We included only peer-reviewed journal articles and accepted full conference proceedings that presented empirical findings on community-based and/or professional fact-checking on social media with relevance to trust, scalability, or impact. We excluded records that focused solely on automated detection, media literacy, or general content moderation when they did not examine human, community-based, or

institutional fact-checking. From an initial yield of 94 unique records, 21 studies were retained after title, abstract, and full-text screening.

The final sample varied by study type (Table 1), region (Table 2), and methodological approach (Table 3). Studies of community-based fact-checking made up the largest share, followed by professional fact-checking studies and comparative studies. More than half were conducted in North America or were U.S.-oriented. Methodologically, over half relied primarily on computational or platform-based analyses, while one-third used experiments. Topics included elections and public affairs, COVID-19 and vaccination, health misinformation, and general political or social misinformation. In addition, 13 of the reviewed studies (61.9%) were conducted in the context of Birdwatch or Community Notes. This reveals how the sample reflects the structure of the emerging field itself: it is currently more focused on crowd-based fact-checking models, U.S.-centered platform environments, and measurable platform dynamics than on professional fact-checking in diverse contexts or on deeper user-oriented outcomes. In terms of platform coverage, Twitter/X was the dominant platform context, especially through studies of Birdwatch and Community Notes. Smaller numbers of studies examined Facebook, WhatsApp, and multi-platform or platform-unspecified contexts. Sample sizes varied considerably across methodological approaches: qualitative studies ranged from small expert interview samples (e.g., $N = 12$) to large survey experiments (e.g., $N = 9,512$), while computational studies analysed datasets ranging from approximately 4,900 Birdwatch notes to over 323,000 Community Notes and more than 2 million ratings. This variation in scale and method is reflected in the coding approach described in the Conceptual Framework section below.

Table 1. *Distribution of reviewed studies by study type ($N = 21$).*

Type of fact-checking studied	N	Percentage (%)
Professional fact-checking only	7	33.3
Community-based fact-checking only	9	42.9
Comparative studies	5	23.8
Total	21	100

Table 2. *Distribution of reviewed studies by region ($N = 21$).*

Region	N	Percentage (%)
North America/U.S.	11	52.4
Asia	4	19.0
Latin America	2	9.5
Europe	1	4.8
Global/transnational	1	4.8
Region not specified	2	9.5
Total	21	100

Note: "U.S.-oriented" studies were coded as North America / U.S. studies described as "U.S. and global" were coded as Global/transnational. Multi-country studies from Argentina, Brazil, Chile, and Colombia were coded as Latin America.

Table 3. Overview of methods in reviewed studies (N = 21).

Methods	N	Percentage (%)
Computational/behavioral data analysis	12	57.1
Experiments	7	33.3
Survey/questionnaires	6	28.6
Interviews/focus groups	3	14.3
Expert assessment	1	4.8

Note: Percentages exceed 100% because some studies used more than one method.

Conceptual framework

The included studies varied substantially in research design, platform context, and outcome measures. Methodological heterogeneity of this kind is a common challenge in systematic reviews of complex interventions and makes it difficult to compare findings using a single standardized metric (Pigott & Shepperd, 2013). We therefore developed a simple conceptual framework, illustrated in Figure 2, to support a more systematic cross-study analysis. The framework was informed by recurring factors in the reviewed literature that either support or limit fact-checking effectiveness. It treats effective fact-checking as dependent on the alignment of three dimensions: trust, scalability, and impact. Trust refers to users' perceived credibility of fact-checking outputs as credible, transparent, and politically neutral, and whether they are willing to accept corrections as legitimate (Brandtzaeg & Følstad, 2017; Lewandowsky et al., 2017). Scalability refers to resources, processing capacity, and dissemination reach (Allen et al., 2021; Godel et al., 2021). Impact refers to measurable effects on misinformation beliefs, sharing behavior, engagement, and deletion (Porter & Wood, 2021). These dimensions were applied across all included studies, regardless of whether they examined professional, community-based, or comparative approaches.

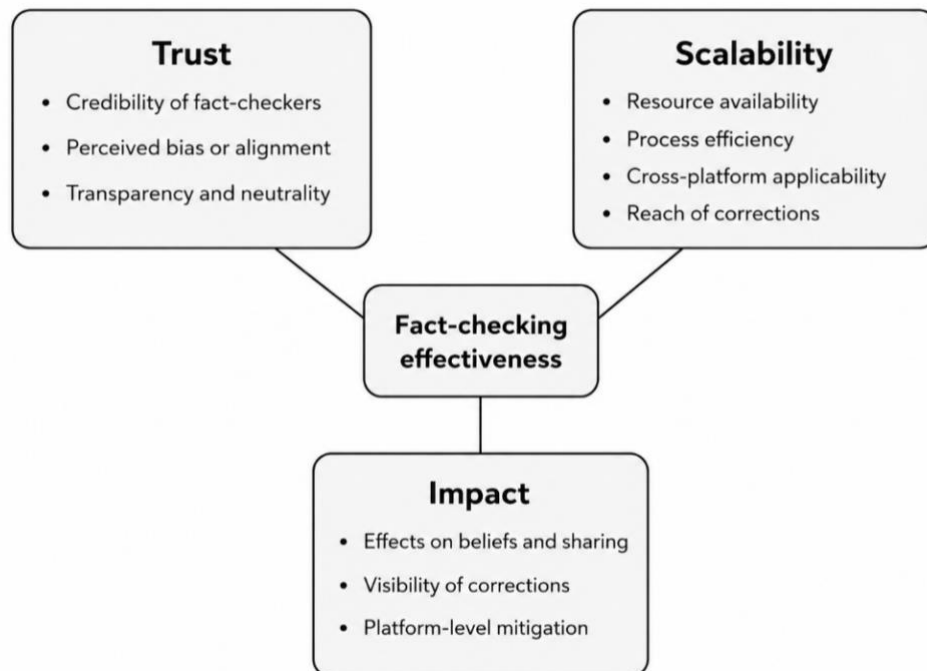


Figure 2. Conceptual framework for evaluating and comparing fact-checking approaches across trust, scalability, and impact.

Each reviewed study was rated as low, moderate, or high on trust, scalability, and impact (see Appendix C for the rating criteria). To account for heterogeneity across the reviewed studies, the ratings were developed through a collaborative, consensus-based process. Three researchers first read the included studies independently and then discussed each study together until agreement was reached on the final ratings. We did not calculate formal intercoder reliability because the reviewed studies were too heterogeneous for a standardized reliability test to be meaningful. As O'Connor and Joffe (2020) note, formal intercoder reliability is not universally appropriate in qualitative or interpretive research, particularly when the task involves context-sensitive judgment rather than uniform coding of comparable material.

Finally, this review has several limitations. The literature is heavily concentrated on Twitter/X, especially in studies of Community Notes and Birdwatch, which limits generalizability to other platforms. The studies are methodologically heterogeneous, and many focus on short-term engagement or belief outcomes rather than long-term behavioral effects. Not all studies fully capture the growing role of AI-generated misinformation, which may place even greater pressure on future fact-checking systems (Bashardoust et al., 2024; Schroeder et al., 2026).

Bibliography

- Allen, J., Arechar, A. A., Pennycook, G., & Rand, D. G. (2021). Scaling up fact-checking using the wisdom of crowds. *Science Advances*, 7(22), eabf4393. <https://doi.org/10.1126/sciadv.abf4393>
- Allen, J., Martel, C., & Rand, D. G. (2022). 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 S. Barbosa, C. Lampe, C. Appert, D. A. Shamma, S. Drucker, J. Williamson, & K. Yatani (Eds.), *CHI '22: Proceedings of the Conference on Human Factors in Computing Systems*. Association for Computing Machinery. <https://doi.org/10.1145/3491102.3502040>
- Allen, M. R., Desai, N., Namazi, A., et al. (2024). Characteristics of X (formerly Twitter) Community Notes addressing COVID-19 vaccine misinformation. *JAMA*, 331(19), 1670–1672. <https://doi.org/10.1001/jama.2024.4800>
- Aruguete, N., Bachmann, I., Calvo, E., Valenzuela, S., & Ventura, T. (2025). Truth be told: How “true” and “false” labels influence user engagement with fact-checks. *New Media & Society*, 27(3), 1443–1464. <https://doi.org/10.1177/14614448231193709>
- Aruguete, N., Batista, F., Calvo, E., Guizzo-Altube, M., Scartaschini, C., & Ventura, T. (2024). Framing fact-checks as a “confirmation” increases engagement with corrections of misinformation: A four-country study. *Scientific Reports*, 14, Article 3201. <https://doi.org/10.1038/s41598-024-53337-0>
- Bartsch, A., Neuberger, C., Stark, B., Karnowski, V., Maurer, M., Pentzold, C., Quandt, T., Quiring, O., & Schemer, C. (2025). Epistemic authority in the digital public sphere: An integrative conceptual framework and research agenda. *Communication Theory*, 35(1), 37–50. <https://doi.org/10.1093/ct/qtae020>
- Bashardoust, A., Feuerriegel, S., & Shrestha, Y. R. (2024). Comparing the willingness to share for human-generated vs. AI-generated fake news. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2), Article 489. <https://doi.org/10.1145/3687028>
- Brandtzaeg, P. B., & Følstad, A. (2017). Trust and distrust in online fact-checking services. *Communications of the ACM*, 60(9), 65–71. <https://doi.org/10.1145/3122803>
- Chuai, Y., Tian, H., Pröllochs, N., & Lenzini, G. (2024). Did the roll-out of Community Notes reduce engagement with misinformation on X/Twitter? *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2), Article 428. <https://doi.org/10.1145/3686967>

- Cotter, K. (2021). "Shadowbanning is not a thing": Black box gaslighting and the power to independently know and credibly critique algorithms. *Information, Communication & Society*, 26(6), 1226–1243. <https://doi.org/10.1080/1369118X.2021.1994624>
- Drolsbach, C. P., & Pröllochs, N. (2022). Diffusion of community fact-checked misinformation on Twitter. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2), Article 428. <https://doi.org/10.1145/3610058>
- Drolsbach, C. P., Solovev, K., & Pröllochs, N. (2024). Community Notes increase trust in fact-checking on social media. *PNAS Nexus*, 3(7), pgae217. <https://doi.org/10.1093/pnasnexus/pgae217>
- Ejaz, W., Altay, S., & Ittefaq, M. (2025). How effective are fact-checks in Pakistan and who engages with them? *Information, Communication & Society*, 28(14), 2439–2461. <https://doi.org/10.1080/1369118X.2024.2445636>
- Flamini, D. (2019, July 3). *Most Republicans don't trust fact-checkers, and most Americans don't trust the media*. Poynter. <https://www.poynter.org/ifcn/2019/most-republicans-dont-trust-fact-checkers-and-most-americans-dont-trust-the-media/>
- Godel, W., Sanderson, Z., Aslett, K., Nagler, J., Bonneau, R., Persily, N., & Tucker, J. A. (2021). Moderating with the mob: Evaluating the efficacy of real-time crowdsourced fact-checking. *Journal of Online Trust and Safety*, 1(1). <https://doi.org/10.54501/jots.v1i1.15>
- He, B., Hu, Y., Lee, Y.-C., Oh, S., Verma, G., & Kumar, S. (2025). A survey on the role of crowds in combating online misinformation: Annotators, evaluators, and creators. *ACM Transactions on Knowledge Discovery from Data*, 19(1), Article 10. <https://doi.org/10.1145/3694980>
- Kuuse, S. F., Kangur, U., Chakraborty, R., & Sharma, R. (2025). Crowdsourced fact-checking or biased commentary? Analyzing political bias in Twitter's Community Notes. In *WWW '25: Companion Proceedings of the ACM on Web Conference 2025* (pp. 2661–2669). Association for Computing Machinery. <https://doi.org/10.1145/3701716.3717533>
- Kyriakidou, M., Cushion, S., Hughes, C., & Morani, M. (2022). Questioning fact-checking in the fight against disinformation: An audience perspective. *Journalism Studies*, 23(15–16), 2123–2139. <https://doi.org/10.1080/17512786.2022.2097118>
- La Barbera, D., Maddalena, E., Soprano, M., Roitero, K., Demartini, G., Ceolin, D., Spina, D., & Mizzaro, S. (2024). Crowdsourced fact-checking: Does it actually work? *Information Processing & Management*, 61(5), Article 103792. <https://doi.org/10.1016/j.ipm.2024.103792>
- Lavilles, R. Q., Naga, J. F., Tinam-isan, M. A. C., Perez, J. E., & Palad, E. B. B. (2023). Information verification practices and perception of social media users on fact-checking services. *Journal of Information Science Theory and Practice*, 11(1), 1–13. <https://doi.org/10.1633/JISTaP.2023.11.1.1>
- Lee, S., Xiong, A., Seo, H., & Lee, D. (2023). "Fact-checking" fact checkers: A data-driven approach. *Harvard Kennedy School (HKS) Misinformation Review*, 4(5). <https://doi.org/10.37016/mr-2020-126>
- Lewandowsky, S. (2025). Free speech, fact checking, and the right to accurate information. *Science*, 387(6632), 123–127. <https://doi.org/10.1126/science.adv4632>
- Lewandowsky, S., Ecker, U. K. H., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition*, 6(4), 353–369. <https://doi.org/10.1016/j.jarmac.2017.07.008>
- Liu, X., Qi, L., Wang, L., & Metzger, M. J. (2025). Checking the fact-checkers: The role of source type, perceived credibility, and individual differences in fact-checking effectiveness. *Communication Research*, 52(6), 719–746. <https://doi.org/10.1177/00936502231206419>
- Meta. (2025, March). Testing begins for Community Notes on Facebook, Instagram and Threads. *Meta Newsroom*. <https://about.fb.com/news/2025/03/testing-begins-community-notes-facebook-instagram-threads/>

- Metzler, H., & Garcia, D. (2023). Social drivers and algorithmic mechanisms on digital media. *Perspectives on Psychological Science*, 19(5), 735–748.
<https://doi.org/10.1177/17456916231185057>
- Nyhan, B. (2021). Why the backfire effect does not explain the durability of political misperceptions. *Proceedings of the National Academy of Sciences*, 118(15), e1912440117.
<https://doi.org/10.1073/pnas.1912440117>
- O'Connor, C., & Joffe, H. (2020). Intercoder reliability in qualitative research: Debates and practical guidelines. *International Journal of Qualitative Methods*, 19.
<https://doi.org/10.1177/1609406919899220>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A... & Moher, D. (2020). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. <https://doi.org/10.1136/bmj.n71>
- Park, C. S. (2023). Why people rely on fact-checkers? Testing theses of “perceived severity of fake news” and “disappointment in news media.” *Journalism Studies*, 25(10), 1–18.
<https://doi.org/10.1080/1461670X.2023.2289878>
- Pigott, T. D., & Shepperd, S. (2013). Identifying, documenting, and examining heterogeneity in systematic reviews of complex interventions. *Journal of Clinical Epidemiology*, 66(11), 1244–1250. <https://doi.org/10.1016/j.jclinepi.2013.06.013>
- Pilarski, M., Solovev, K. O., & Pröllochs, N. (2024). Community Notes vs. Snopes: How the crowd selects fact-checking targets on social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 18(1), 1262–1275. <https://doi.org/10.1609/icwsm.v18i1.31387>
- Porter, E., & Wood, T. J. (2021). The global effectiveness of fact-checking: Evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom. *Proceedings of the National Academy of Sciences*, 118(37), e2104235118.
<https://doi.org/10.1073/pnas.2104235118>
- Pretus, C., Gil-Buitrago, H., Cisma, I., Hendricks, R. C., & Lizarazo-Villarreal, D. (2024). Scaling crowdsourcing interventions to combat partisan misinformation. *Advances in Psychology*, 2, e85592. <https://doi.org/10.56296/aip00018>
- Primig, F. (2022). The influence of media trust and normative role expectations on the credibility of fact checkers. *Journalism Practice*, 18(5), 1137–1157.
<https://doi.org/10.1080/17512786.2022.2080102>
- Pröllochs, N. (2022). Community-based fact-checking on Twitter’s Birdwatch platform. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1), 794–805.
<https://doi.org/10.1609/icwsm.v16i1.19335>
- Roozenbeek, J., Remshard, M., & Kyrychenko, Y. (2024). Beyond the headlines: On the efficacy and effectiveness of misinformation interventions. *Advances in Psychology*, 2, e24569.
<https://doi.org/10.56296/aip00019>
- Saeed, M., Traub, N., Nicolas, M., Demartini, G., & Papotti, P. (2022). Crowdsourced fact-checking at Twitter: How does the crowd compare with experts? In *CIKM '22: Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (pp. 1736–1746). Association for Computing Machinery. <https://doi.org/10.1145/3511808.3557279>
- Schroeder, D. T., Cha, M., Baronchelli, A., Bostrom, N., Christakis, N. A., Garcia, D., Goldenberg, A., Kyrychenko, Y., Leyton-Brown, K., Lutz, N., Marcus, G., Menczer, F., Pennycook, G., Rand, D. G., Schweitzer, F., Summerfield, C., Tang, A., Van Bavel, J., van der Linden, S., Song, D., & Kunst, J. R. (2026). How malicious AI swarms can threaten democracy. *Science*, 391(6783), 354–357.
<https://doi.org/10.1126/science.adz1697>

- Seelam, A., Paul Choudhury, A., Liu, C., Goay, M., Bali, K., & Vashistha, A. (2024). “Fact-checks are for the Top 0.1%”: Examining reach, awareness, and relevance of fact-checking in rural India. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1), Article 56. <https://doi.org/10.1145/3637333>
- Swire-Thompson, B., Miklaucic, N., Wihbey, J. P., Lazer, D., & DeGutis, J. (2022). The backfire effect after correcting misinformation is strongly associated with reliability. *Journal of Experimental Psychology: General*, 151(7), 1655–1665. <https://doi.org/10.1037/xge0001131>
- Tan, N. (2020). Electoral management of digital campaigns and disinformation in East and Southeast Asia. *Election Law Journal: Rules, Politics, and Policy*, 19(2), 214–239. <https://doi.org/10.1089/elj.2019.0599>
- Zhao, A., & Naaman, M. (2023). Insights from a comparative study on the variety, velocity, veracity, and viability of crowdsourced and professional fact-checking services. *Journal of Online Trust and Safety*, 2(1). <https://doi.org/10.54501/jots.v2i1.118>
- Zhou, Y., Hou, J., Gao, Y., & Chen, P.-Y. (2025). How does crowdsourced fact-checking approach tackling misinformation affect audience engagement? Evidence from Twitter’s Community Notes program. In T. X. Bui (Ed.), *Proceedings of the 58th Hawaii International Conference on System Sciences* (pp. 6554–6561). University of Hawaii. <https://doi.org/10.24251/HICSS.2025.783>

Funding

This project was funded by the Norwegian Research Council (grant agreement no. 352482) and the AI-democracy programme at the University of Oslo. They had no role in the design, interpretation, or reporting of the research.

Competing interests

The authors declare no competing interests.

Ethics

This review protocol was exempt from ethics approval.

Copyright

This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are properly credited.

Data availability

All materials needed to replicate this study are available via the Harvard Dataverse: <https://doi.org/10.7910/DVN/JMGKFL>

Appendix A: Findings

Tables A1, A2, and A3 synthesize the findings from 21 peer-reviewed studies examining fact-checking interventions on social media, focusing on three dimensions: trust, scalability, and impact. The tables also report platform context, method, region, and, where available, sample size or corpus size. Our analysis shows how methodological choices shape the assessment of these dimensions. Studies of professional fact-checking primarily examine individual-level trust and attitudinal outcomes using experiments, surveys, and interviews, whereas studies of community-based fact-checking more often evaluate scalability and systemic impact through platform-level, data-intensive analyses. The studies are geographically diverse, covering North America, Latin America, Europe, and Asia, but the evidence base remains concentrated in U.S.-oriented X/Twitter research. Platforms examined include X/Twitter, Facebook, WhatsApp, online newspapers, community radio, and professional fact-checking sites, with X/Twitter hosting the most prominent community-based initiative, Community Notes (formerly Birdwatch). Table A1 below sums up the results from the reviewed studies on professional fact-checking. Table A2 shows reviewed studies on community-based fact-checking, while Table A3 shows the reviewed comparative studies.

Table A1. *Studies on professional fact-checking (n = 7).*

Authors	Trust	Scalability	Impact	Platform and method
Aruguete et al. (2024)	N/A	N/A	Moderate; users more likely to like and engage with confirmatory fact-checks aligned with own beliefs.	Facebook. Argentina, Brazil, Chile, and Colombia. Questionnaire-based experiment (N = 9,512). Participants exposed to Facebook-style posts about COVID-19 vaccines and other topics, framed as confirmational (e.g., “True:”) or refutational (e.g., “False:”) fact-checks.
Aruguete et al. (2025)	N/A	N/A	Moderate; users more likely to share confirmatory factchecks that align with own beliefs.	Twitter. Chequeado (leading Argentine fact-checking organization). Argentina. Experimental manipulation of tweet labels; analysis of user engagement behavior (N = 2,041), combined with observation data on Twitter.
Ejaz et al. (2025)	Moderate; the effectiveness and the engagement with both the fact-checks and mis-information where moderate by trust.	Moderate; liking or sharing correction on social media higher among those concerned with mis-information.	Moderate; fact-checking most effective among participants who trust the news the most, and least effective among those trusting social media.	Facebook, Twitter, online newspapers. Pakistan. Questionnaire-based experiment (N participants = 302, N observations = 1,208) investigating the effectiveness of fact-checking, and the determinants of engagement with fact-checks and misinformation.

Authors	Trust	Scalability	Impact	Platform and method
Freiling et al. (2023)	N/A	N/A	Moderate; users' motivational and affective states limit the impact of fact-checking.	Facebook. U.S. Experimental design using real-world misinformation/fact-check content in a politically charged public health context (COVID-19) (N = 719).
Lavilles et al. (2023)	Low: Only 25–27% expressed explicit trust in Rappler and VeraFiles (Those two are the focus in the study, however, two other fact-checking initiatives with higher trust score are mentioned).	Low; only 22% had used Rappler or VeraFiles for verification. Low capacity to respond to all mis-information.	Low; fact-checks with limited behavioral impact. Some users reported fact-checking to improve own verification behavior, others prefer doing own checks.	Facebook, Rappler, VeraFiles. Philippines. Mixed methods: FGDs, focus group interviews (n = 121), survey (n = 685). Filipino social media users in 2020.
Park (2023)	Moderate; perceived severity of fake news positively associated with trust in fact-checking.	N/A	Moderate; perceived severity of fake news and, to some extent dissatisfaction with news media, positively associated with use of fact checkers.	Social media platforms. U.S. Questionnaire study (N = 2,350).

Authors	Trust	Scalability	Impact	Platform and method
Seelam et al. (2024)	Low-moderate; credible by design, but awareness and engagement are limited. Quote: "Fact-checks are for the top 0.1%... educated population[s] at least have the know-how and agency..."	Low; major bottlenecks including financial constraints, limited manpower, and technological gaps in hyperlocal outreach.	Low; fact-checks with some impact via outreach efforts (e.g., stringers, local language materials), but limited by low media literacy, lack of awareness, and infrastructural gaps.	WhatsApp, Facebook, community radio, traditional media. India. Interviews with 12 representatives from fact-checking organizations (e.g., editors, fact checkers, founders).

References (Table A1)

- Aruguete, N., Batista, F., Calvo, E., Guizzo-Altube, M., Scartaschini, C., & Ventura, T. (2024). Framing fact-checks as a "confirmation" increases engagement with corrections of misinformation: A four-country study. *Scientific Reports*, 14, Article 3201. <https://doi.org/10.1038/s41598-024-53337-0>
- Aruguete, N., Bachmann, I., Calvo, E., Valenzuela, S., & Ventura, T. (2025). Truth be told: How "true" and "false" labels influence user engagement with fact-checks. *New Media & Society*, 27(3), 1443–1464. <https://doi.org/10.1177/14614448231193709>
- Ejaz, W., Altay, S., & Ittefaq, M. (2025). How effective are fact-checks in Pakistan and who engages with them? *Information, Communication & Society*, 28(14), 2439–2461. <https://doi.org/10.1080/1369118X.2024.2445636>
- Freiling, I., Krause, N. M., Scheufele, D. A., & Brossard, D. (2023). Believing and sharing misinformation, fact-checks, and accurate information on social media: The role of anxiety during COVID-19. *New Media & Society*, 25(1), 141–162. <https://doi.org/10.1177/14614448211011451>
- Lavilles, R. Q., Naga, J. F., Tinam-isan, M. A. C., Perez, J. E., & Palad, E. B. B. (2023). Information verification practices and perception of social media users on fact-checking services. *Journal of Information Science Theory and Practice*, 11(1), 1–13. <https://doi.org/10.1633/JISTaP.2023.11.1.1>
- Park, C. S. (2023). Why people rely on fact-checkers? Testing theses of "perceived severity of fake news" and "disappointment in news media." *Journalism Studies*, 25(1), 1–18. <https://doi.org/10.1080/1461670X.2023.2289878>
- Seelam, A., Paul Choudhury, A., Liu, C., Goay, M., Bali, K., & Vashistha, A. (2024). "Fact-checks are for the Top 0.1%": Examining reach, awareness, and relevance of fact-checking in rural India. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1), Article 56. <https://doi.org/10.1145/3637333>

Table A2. *Studies on community-based fact-checking (n = 9).*

Authors	Trust	Scalability	Impact	Platform and method
Allen et al. (2022)	N/A	N/A	Moderate; partisan bias may weaken corrective effects, limits visibility and diminishes real-world influence of Birdwatch notes.	X-platform. U.S.-oriented. Analysis of Twitter/Birdwatch data on tweets, notes (>4,9K), and ratings (>28K). Concerning how political partisanship influences users' behavior within Twitter's experimental fact-checking Birdwatch.
Allen et al. (2024)	N/A	High; 97% of notes found to be scientifically accurate. Posts with the studied notes reached over 200 million views. Number of COVID-specific notes was still low relative to the problem's size.	N/A	X-platform. U.S.-oriented. Expert assessment of 205 Community Notes on COVID-19 vaccination. Notes gathered from open data.
Chuai et al. (2024)	N/A	Moderate; substantial participation in Community Notes (>64K contributors in study sample).	Low; no significant effect of Community Notes, assumed due to timing delays. (This in contrast to Twitter/X's own studies during pilot, where an effect of notes was found—likely due to timing issues).	X-platform. U.S. and global. Analysis of Notes (>68K), ratings (2,3M), and enrolled contributors (>64K) during U.S. and global implementation of the Twitter/X Community Note feature (2021–2023).

Authors	Trust	Scalability	Impact	Platform and method
Drolsbach & Pröllochs (2022)	N/A	High; Birdwatch found to leverage large numbers of ordinary users to identify, annotate, and contextualize misleading content in real time.	Moderate; Birdwatch notes associated with misleading posts receiving fewer retweets than non-misleading posts, indicating reduced diffusion.	X-platform. U.S.-oriented. Behavioral and observational data analysis of Birdwatch.
Drolsbach et al. (2024)	Moderate; Community Notes were rated as more trustworthy than simple mis-information flags. Skepticism among politically misaligned users.	N/A	Moderate; Community Notes increase participants' ability to correctly identify misleading posts compared to no fact-checking and simple flags. Variation across partisan lines.	X-platform. U.S. questionnaire-based experiment (N = 1,810) involving a sample of 36 misleading and non-misleading social media posts. Assessment of users' trust in different types of fact-checking interventions
Kuuse et al. (2025)	N/A	Moderate; Community Notes demonstrate potential for wide coverage and volume processing, yet qualitative consistency varies across ideological and contextual dimensions.	N/A	X-platform. English language. Large-scale computational content analysis of 323,382 Community Notes, collected between Jan. 2021 and Nov. 2023, using bias scoring and sentiment analysis.

Authors	Trust	Scalability	Impact	Platform and method
La Barbera et al. (2024)	N/A	Moderate; scales well via crowd-sourcing, but high consensus threshold limits broad visibility.	N/A	X-platform. U.S.-oriented. NLP-based lexical, sentiment, and temporal analysis; source bias via media monitors. Challenge with ecological validity (what is tested, is whether anyone can fact-check in view of a selection of parameters).
Pröllochs (2022)	Moderate; struggles with consensus. Skepticism about bias but recognizes transparency mechanisms.	N/A	Moderate; Community Notes reduce tweet visibility and engagement but may be unevenly distributed due to polarization.	X-platform. (Birdwatch on then Twitter). U.S.-oriented. Behavioral data from 11,802 notes analyzed.
Zhou et al. (2025)	N/A	N/A	Moderate; Community Notes can decrease audience engagement with labeled authors long-term but boost short-term engagement.	X-platform. Region not specified. Analysis of Community Notes, investigating audience engagement with authors of misinformation.

References (Table A2)

- Allen, J., Martel, C., & Rand, D. G. (2022). 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 S. Barbosa, C. Lampe, C. Appert, D. A. Shamma, S. Drucker, J. Williamson, & K. Yatani (Eds.), *CHI '22: Proceedings of the conference on human factors in computing systems*. Association for Computing Machinery. <https://doi.org/10.1145/3491102.3502040>
- Allen, M. R., Desai, N., Namazi, A., et al. (2024). Characteristics of X (formerly Twitter) Community Notes addressing COVID-19 vaccine misinformation. *JAMA*, 331(19), 1670–1672. <https://doi.org/10.1001/jama.2024.4800>
- Chuai, Y., Tian, H., Pröllochs, N., & Lenzini, G. (2024a). Did the roll-out of Community Notes reduce engagement with misinformation on X/Twitter? *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2), Article 428. <https://doi.org/10.1145/3686967>
- Drolsbach, C., & Pröllochs, N. (2022). Diffusion of community fact-checked misinformation on Twitter. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2), Article 428. <https://doi.org/10.1145/3610058>
- Drolsbach, C. P., Solovev, K., & Pröllochs, N. (2024). Community Notes increase trust in fact-checking on social media. *PNAS Nexus*, 3(7), pgae217. <https://doi.org/10.1093/pnasnexus/pgae217>

- Kuuse, S. F., Kangur, U., Chakraborty, R., & Sharma, R. (2025). Crowdsourced fact-checking or biased commentary? Analyzing political bias in Twitter's Community Notes. In WWW '25: *Companion Proceedings of the ACM on Web Conference 2025* (pp. 2661–2669). Association for Computing Machinery. <https://doi.org/10.1145/3701716.3717533>
- La Barbera, D., Maddalena, E., Soprano, M., Roitero, K., Demartini, G., Ceolin, D., Spina, D., & Mizzaro, S. (2024). Crowdsourced fact-checking: Does it actually work? *Information Processing & Management*, 61(5), Article 103792. <https://doi.org/10.1016/j.ipm.2024.103792>
- Pröllochs, N. (2022). Community-based fact-checking on Twitter's Birdwatch platform. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1), 794–805. <https://doi.org/10.1609/icwsm.v16i1.19335>
- Zhou, Y., Hou, J., Gao, Y., & Chen, P.-Y. (2025). How does crowdsourced fact-checking approach tackling misinformation affect audience engagement? Evidence from Twitter's Community Notes program. In T. X. Bui (Ed.), *Proceedings of the 58th Hawaii International Conference on System Sciences* (pp. 6554–6561). University of Hawaii. <https://doi.org/10.24251/HICSS.2025.783>

Table A3. Comparative studies of professional and community-based fact-checking (n = 5).

Authors	Trust	Scalability	Impact	Platform and method
Liu et al. (2025)	Moderate; professional fact checkers perceived as significantly more credible than news outlets and social media, but at same level as community notes and AI-based fact-checking.	N/A	Moderate: reduction in misinformation for all fact-checking, but no significant difference between professional fact checkers and community notes.	Social media, news media, and AI-based fact-checking. U.S. online experiment (N = 859) comparing professional fact checker (FactCheck.org), community notes (Birdwatch), and others, including news media, social media, and AI-based fact-checking.

Authors	Trust	Scalability	Impact	Platform and method
Pilarski et al. (2024)	N/A	Professional Moderate; broader coverage and more diverse target base, less visibility in high-traffic threads may limit reach. Community Moderate; target highly viral and visible posts, but slow response time.	Professional N/A Community N/A	X-platform. U.S. Comparative analysis of Community Notes (>25K) and replies with links to professional fact checkers (52K), during Community Notes' pilot phase (Jan. 2021–Oct. 2022).
Primig (2022)	Professional Moderate; (dpa, Correctiv) were perceived as more trustworthy but also perceived by some as state-aligned or propaganda.	Professional Low; real-world reach is limited: many in the sample had not encountered the fact-checks organically and only reacted when shown to them in the study.	Professional Moderate; reduced misinformation among participants who trust media and institutions.	Facebook. Germany. Investigates how German Facebook users assess the credibility of fact-checking sources (dpa, Corrective vs a fictitious "Facebook Community") and messages related to COVID-19 disinformation (N = 123).
	Community Low; "Facebook community" source was rated as least credible.	Community N/A	Community Low; no evidence of change in perception of community-checked content.	

Authors	Trust	Scalability	Impact	Platform and method
Saeed et al. 2022)	Professional High; experts rely on relatively small set of high-quality sources to verify claims.	Professional Low; scale globally but are slower and less responsive to novel mis-information.	Professional Moderate; have longer-term and institutional influence.	Birdwatch (Twitter/X): U.S. Comparing crowdsourced fact-checking (Birdwatch) with expert fact-checking and using computational methods. ClaimReview checks contains 76,769 links covering 73 domains of fact-checking groups and journalists.
	Community Moderate; cites diverse sources, most evaluated as credible and useful, but are more vulnerable to partisan bias, credibility challenge identified as malign users may game the algorithm.	Community High; rapid crowd response (~10× faster than experts).	Community Moderate; high potential impact due to speed and breadth, but - vulnerable to partisanship.	

Authors	Trust	Scalability	Impact	Platform and method
Zhao & Naaman (2023)	Professional High; crowdsourced fact checkers cite professionals heavily (454/897 cases). Dominates expertise-dependent topics (medicine and health, international affairs).	Professional Low; focuses on complex, low-volume topics. Slower response (>24hr median for new claims).	Professional Moderate; critical impact in specialized domains (COVID-19). Enables crowdsourced scale (50% citations).	A variety of platforms. Taiwan. Compares a fact-checking community (Cofacts) with two professional sites (MyGoPen and Taiwan FactCheck Center), on “variety” (breadth of coverage/topic coverage), “velocity” (speed in answering new requests), veracity (accuracy, reliability, and objectivity), and viability (persuasiveness and effectiveness of fact-checks).
	Community N/A	Community High; higher volume (e.g., 11,747 vs. 418 health articles). 57% faster response for new claims (faster in 754/879 cases).	Community N/A	

References (Table A3)

- Liu, X., Qi, L., Wang, L., & Metzger, M. J. (2025). Checking the fact-checkers: The role of source type, perceived credibility, and individual differences in fact-checking effectiveness. *Communication Research*, 52(6), 719–746. <https://doi.org/10.1177/00936502231206419>
- Pilarski, M., Solovev, K. O., & Pröllochs, N. (2024). Community Notes vs. snoping: How the crowd selects fact-checking targets on social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 18(1), 1262–1275. <https://doi.org/10.1609/icwsm.v18i1.31387>
- Primig, F. (2022). The influence of media trust and normative role expectations on the credibility of fact checkers. *Journalism Practice*, 18(5), 1137–1157. <https://doi.org/10.1080/17512786.2022.2080102>
- Saeed, M., Traub, N., Nicolas, M., Demartini, G., & Papotti, P. (2022). Crowdsourced fact-checking at Twitter: How does the crowd compare with experts? In *CIKM '22: Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (pp. 1736–1746). Association for Computing Machinery. <https://doi.org/10.1145/3511808.3557279>

Zhao, A., & Naaman, M. (2023). Insights from a comparative study on the variety, velocity, veracity, and viability of crowdsourced and professional fact-checking services. *Journal of Online Trust and Safety*, 2(1). <https://doi.org/10.54501/jots.v2i1.118>

Appendix B: Summary of findings

Table B1 summarizes the main pattern in the review. Professional and community-based fact-checking do not perform equally across trust, scalability, and impact. Professional fact-checking tends to perform somewhat better on trust but is more limited in speed and reach. Community-based fact-checking performs better on scalability, but impact is constrained by note visibility, consensus thresholds, and uneven user acceptance. Both models therefore show clear strengths, but neither performs strongly across all three dimensions at once. The table should not be read as showing that one model is better overall. Rather, it shows that current fact-checking systems solve different parts of the misinformation problem while remaining limited in others.

Table B1. *Comparative performance of professional and community-based fact-checking across three dimensions (N = 21).*

Dimension	Professional fact-checking	Community-based fact-checking
Trust	Moderate — credibility depends on institutional reputation and prior media trust; perceived as biased by politically disaffected audiences	Moderate — participatory transparency builds some trust, but ideological skew and inconsistent note quality create scepticism
Scalability	Low–Moderate — manual workflows, limited staff, ~100 verified claims per day; slow to respond to viral content	High — distributed labor enables near real-time annotation; up to 10× faster response than professional checkers
Impact	Moderate — reduces false beliefs, especially among high-trust users, but limited organic reach constrains population-level effects	Moderate — reduces retweet volume and encourages content deletion, but consensus thresholds and low note visibility limit broader corrective effects

Appendix C: Eligibility, search, and coding protocol

Review design

This study used a PRISMA-guided systematic review design to identify and synthesize empirical studies of human-driven professional and community-based fact-checking on social media.

Databases and search period

Searches were conducted in ACM Digital Library, Web of Science, Scopus, and JSTOR. Eligible publications were published from 2022 through 2025.

Search strategy

The search strategy comprised three conceptual groups: misinformation, social media platforms, and fact-checking or verification. Terms within each group were joined with OR, while the three groups were joined with AND.

("misinformation" OR "disinformation") AND ("social media" OR "platform*" OR "Twitter" OR "X" OR "Facebook" OR "WhatsApp") AND ("fact-checking" OR "fact checking" OR "fact-check*" OR "misinformation correction" OR "Community Notes" OR "Birdwatch" OR "community fact-checking" OR "professional fact-checking" OR "crowdsourced verification" OR "peer fact-checking" OR "expert verification")

Trust, scalability, and impact were analytical categories and were not required search terms.

Inclusion criteria

1. Peer-reviewed journal article or accepted full conference proceeding.
2. Empirical study.
3. Examined professional and/or community-based human fact-checking.
4. Fact-checking occurred on or was directly connected to social media.
5. Reported evidence relevant to trust, scalability, or impact.
6. Published from 2022 through 2025.

Exclusion criteria

1. Not peer-reviewed.
2. Not empirical.
3. Did not examine professional or community-based human fact-checking.
4. Did not concern social media.
5. Focused only on automated detection, media literacy, or general content moderation without examining human or institutional fact-checking.
6. Examined crowd accuracy judgments rather than deployed platform fact-checking with public annotations or visible contributor participation.

Screening

Records were screened by title and abstract and then by full text against the eligibility criteria. The search yielded 94 unique records. Twenty-five full texts were assessed for eligibility, four were excluded at the full-text stage, and 21 studies were included in the review.

The aggregate screening totals reported here should be read together with the original screening log, which provides the record-level inclusion and exclusion decisions.

Coding and synthesis

Three researchers read the included studies independently. Their interpretations were discussed collectively and refined through comparison until agreement was reached. Formal intercoder reliability was not calculated because the evidence was methodologically heterogeneous and the ratings required context-sensitive judgment.

Each study was assessed for trust, scalability, and impact. The ratings were interpretive summaries of the strength, clarity, and consistency of the evidence reported within each study; they were not standardized quantitative effect measures or assessments of overall study quality. Table C1 defines the thresholds used in the synthesis.

Table C1. Coding framework for trust, scalability, and impact.

Rating	General threshold	Evidence characteristics
High	Strong, consistent evidence across contexts; clear patterns of effectiveness, reach, and/or credibility	Clear author claims, multiple sources, and/or converging data (e.g., behavioral, perceptual, platform analytics). Effects generalize well across groups and/or settings.
Moderate	Conditional or mixed evidence; effectiveness depends on context, user group, and/or platform	Qualifiers present (“some”, “often,” “under certain conditions”); partial and/or domain-specific effects; trust/scalability/impact varies by audience and/or content.
Low	Weak, inconsistent, and/or negative evidence; minimal reach, trust, and/or behavioral effect	Lack of measurable outcomes; evidence of failure, disengagement, low trust, and/or structural bottlenecks; authors report limitations and/or ineffectiveness.
N/A	Insufficient evidence to assess the dimension	The study did not report evidence relevant to the dimension. N/A does not indicate weak or low performance.