THE TRUST-CONSENSUS PARADOX: WHY DECENTRALIZED FACT-CHECKING FACES CHALLENGES ON POLARIZING TOPICS

By: Valeria de la Fuente, Nathan Doctor, and Alexander Hohlfeld

ISD analyzed one year of public data from X's Community Notes model to assess its effectiveness and identify key limitations relevant for Meta's own implementation. Our findings reveal a critical tension at the center of this model: the same structural features that make the system transparent, credible and collaborative also limit its effectiveness in addressing high-stakes, harmful or polarizing content. While the model's bottom-up approach fosters trust, encourages diverse perspectives and promotes user engagement, it is often slow to respond to rapidly spreading misinformation and lacks the rigor of traditional fact-checking.

Key Findings

- **Community Notes appear to be widely trusted across the political spectrum.** This seems to stem from its collaborative, bottom-up design, transparency and the inclusion of explanatory context which many users find more persuasive than top-down branded fact-checks. Even users who have previously spread misinformation or received corrections from Notes themselves often express support for the program.

- **Polarization limits Notes' effectiveness against controversial but misleading content.** The consensus needed for Community Notes to be effective is often elusive, especially on politically-charged content. Based on an LLM classifier trained to detect political content and a manual review of a sample of 50 helpful Notes, we found roughly half of Community Notes were applied to "soft news" (stories covering culture, lifestyle, or lighter topics rather than politics or current events). By comparison, factually accurate and well-sourced Notes surrounding politically charged events frequently went unpublished, leaving harmful misinformation unchallenged.

- **Trust in Community Notes and their ability to take on controversial content may be inversely related.** If Notes were applied more frequently, trust in the program could be eroded. This paradox of consensus may render the system incapable of targeting controversial content while maintaining broad trust.

- **The program continues to suffer from process-related time lags.** Our research found a median delay of more than 15 hours between the posting of a misleading post and the application of a helpful Community Note. While Notes reduce engagement on misleading content, most views and engagement occur in the first hours of a posts.

- **Community Notes struggle to rise to the occasion during high-volume crises.** In fast-moving events like natural disasters or incidents of political violence, misinformation can spread faster than Community Notes can respond. Our case study on Hurricanes Helene and Milton showed that only 10% of sampled high-engagement false claims received a visible note after an average of 46 hours. This gap poses serious risks in moments when accurate information is most urgently needed.

Conclusions & Recommendations

Our analysis of X's Community Notes program underscores a key reality: misinformation defies a silver bullet solution. While crowd-sourced approaches face notable challenges, they can nevertheless play a meaningful role in addressing the spread of false or misleading content. Based on our findings we offer four key recommendations for platforms considering or currently implementing similar initiatives:

1. Ensure a high level of contributors

The success of any crowd-sourced program depends on broad and sustained user participation. A key challenge facing X's Community Notes program in 2025 has been the stagnation of active contributors. As of November 2025, X reported roughly 1.3 million contributors worldwide. However, over the past two years, program usage has remained relatively flat. This includes both the volume of user ratings (Figure 10) and the number of unique active contributors (Figure 11).

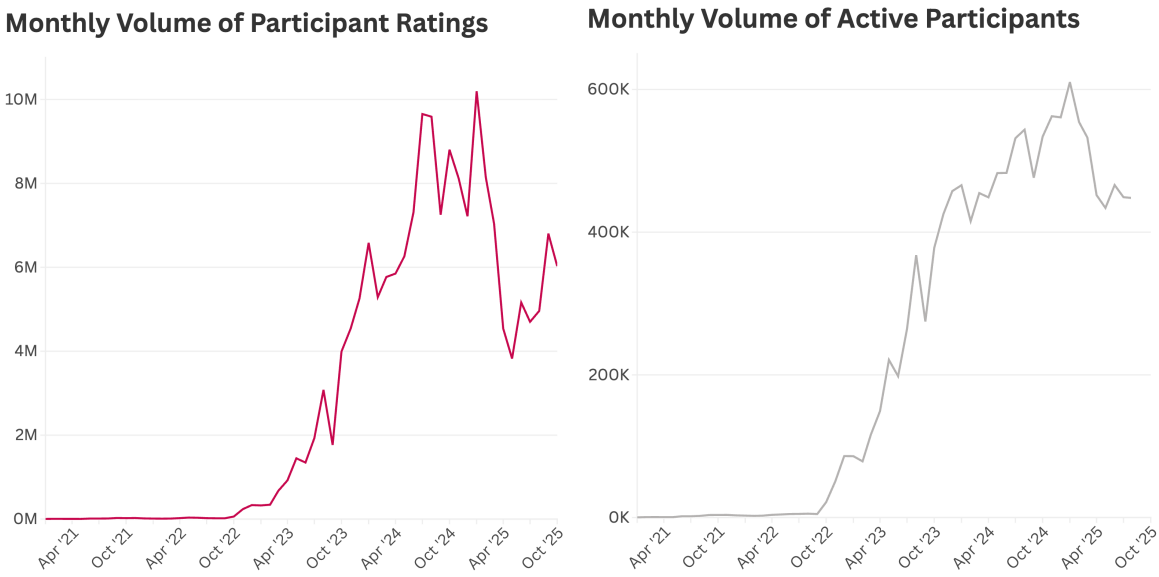

Figure 1 and 2: The volume of participant ratings was calculated by aggregating the total number of ratings applied to Community Notes from January 2021 through October 2025. Active users represent the total number of unique raters each month.

This stagnation likely reflects insufficient incentives for ongoing participation. Many Notes receive few or no ratings, and contributors rarely see tangible impact from their efforts. While such friction is inherent to the program's design – which prioritizes consensus over speed – platforms should work to maximize user engagement and satisfaction.

This should be through a focus on the user experience, such as providing features like achievement systems that reward consistent, high-quality contributions. Such gamification may include user statistics – such as total Notes written, helpfulness ratings, and community impact metrics – while tiered contributor levels can provide tangible benefits and a sense of recognition as users advance.

Comparing contributors across platforms is difficult due to geographic limitations. Both Meta's Community Notes and TikTok's FootNotes are currently only available in the United States, whereas X's program is global. Meta's Chief Information Security Officer claimed in September 2025 that their program would have more than 70,000 active contributors who would have written more than 15,000 Notes from which 6% would have been published. According to some sources, Meta would have admitted more than 250,000 users to contribute to the program. TikTok claimed that 80.000 users would have qualified to contribute to their "FootNotes" program when starting the program in July. While X provides public datasets on the contributors to its Community Notes program, both Meta and TikTok are lacking this transparency, making it impossible to verify or research the level of contribution.

2. Ensure platform transparency & data accessibility

This analysis was only possible due to the transparent nature of the Community Notes and the transparency efforts behind it. X has open-sourced its Community Notes algorithm and provides free access to Notes, ratings, and user enrollment data alongside a comprehensive data guide. This level of openness enables independent scrutiny, fosters trust, and empowers external researchers to identify and help resolve issues. Updates to the algorithm are regularly shared with the public. For example, the public criticism that Community Notes would appear too slow, as also shown by our analysis, led to X recently announcing a new note scoring aggregation as a pilot in November 2025. How this algorithm change will affect the challenges highlighted by our report will be subject to subsequent analysis.

Such openness remains rare across the industry, yet it is essential for building credibility in crowdsourced moderation systems. By contrast, Meta and TikTok have not published the algorithms behind their respective programs, nor do they provide public access to contributor ratings or note data. This lack of transparency not only makes it difficult for researchers and civil society to evaluate the effectiveness or fairness of the systems, but can also negatively affect the public trust — which, as we pointed out, might be one of the strongest aspects of X's Community Notes program.

3. Encourage integrative approaches

As our research has shown, Community Notes are not a silver bullet solution to challenges around mis- and disinformation and should not be seen as such.

The inherent trade-offs and tensions in various fact-checking models underscore the need for integrative frameworks. While Community Notes offer scalability and broader public trust, they often rely on institutional fact-checking, as recent research has demonstrated. A recent study by the Spanish fact-checking organisation Maldita found that fact-checking organisations were the third most used reference for Community Notes and Notes including a link to a fact-checking organisation would have appeared 90 minutes earlier than other Notes on average.

Crowdsourced approaches, such as Community Notes can contribute to increasing the scale and reach of important context while maintaining broader trust. New features like media or link matching, allowing Notes to be automatically associated with posts containing the same images, videos or URLs, further enhance this scalability.

These findings suggest that different approaches can complement each other. Rather than choosing between professional and community-based moderation, platforms should experiment with hybrid models that integrate both. Such models could combine the rigor and expertise of institutional fact-checking with the reach and responsiveness of community driven systems, while maintaining broader trust.

4. Focus on user empowerment

A key reason why Community Notes enjoy high levels of trust lies in their label design. Rather than asserting a top-down version of truth, they frame contributions as "Readers added context they thought people might want to know". This phrasing respects users as autonomous and capable citizens, empowering them to make their own judgements rather than passively accepting authoritative claims. The crowdsourced nature of the system further reinforces a sense of civic recognition and participation. A key lesson from Community Notes for platforms thus lies in the importance of centering user empowerment in platform design.

The full investigation, including detailed figures, methodology, and additional analysis are available at ISD's website: https://isdglobal.org/digital_dispatches/the-trust-consensus-paradox-why-decentralized-fact-checking-faces-challenges-on-polarizing-topics/