Full Fact response to Meta Oversight Board: Community Notes rollout

About Full Fact

Full Fact is a UK charity building a better information environment to restore trust. We believe that facts matter. Our team of independent fact checkers, journalists, technologists and policy experts fact check claims made by politicians, public institutions, and in viral content online; hold public figures and institutions accountable by encouraging corrections; and build world-leading AI tools, used in 40+ countries globally.

Summary

Key recommendations for Community Notes

- Develop or adopt an existing harms framework for Community Notes.
- Community Notes should be piloted, and researchers must be granted access to data for independent evaluation.
- Introduce a fast track system including working with experts, to head off critical threats to the information environment.
- Develop a system to amalgamate unpublished notes on the same topic into single Super Notes, using generative AI trained to get sufficient votes due to clarity, sourcing and balance.

Full Fact favours solutions to misinformation that prioritise freedom of expression, such as providing context, reducing amplification and building information literacy, rather than removing content. **These types of interventions are effective** in containing and reducing belief in harmful misinformation.

It is not clear whether in expanding Community Notes, Meta intends to remove Third-Party Fact-Checking (TPFC), following its actions in the US. **We strongly oppose shutting down TPFC**: fact checking and Community Notes are intertwined and complementary. Fact-checks are the third most-used reference globally when someone proposes a community note on X, and notes with fact checks get published faster.

Birdwatch was in development for two years before Twitter rolled it out as Community Notes in 2022. Yoel Roth, Twitter's former head of trust and safety, was clear that Community Notes was intended to complement rather than replace other misinformation-countering efforts. By **removing TPFC in the US, and replacing it with an untested new system, Meta has left a vast moderation gap**, allowing risks to proliferate around addressing harmful and urgent misinformation in a timely way. This must not be repeated elsewhere. **Community Notes should be piloted, and researchers must be granted access to data for evaluation.**

Every country has areas of highly partisan debate and moments of crisis or vulnerability such as elections or disasters. Identifying harm potential is critical to moderating content in ways that enable freedom of expression while safeguarding against real world harm. Community Notes cannot do this, and fact checkers can. We **challenge the Community Notes model to develop a clearer philosophy of harm**, instead of treating all posts equally.

Answers to the Oversight Board's questions

1. The risks and opportunities of crowd-sourced and community notes-style approaches to content moderation, particularly when it comes to potentially misleading content.

Opportunities

Community notes-style systems enable moderation focused on contextualisation rather than removal, create the potential for more diverse perspectives to be involved in moderating content (and for that to positively affect trust in the system), give the chance for more specialist knowledge and local nuance to appear.

In its early tests, Twitter found that surveys showed people were 20% to 40% less likely to agree with the substance of a potentially misleading Tweet after reading a note about it, compared to those who saw a Tweet without a note. Other research suggests that when a note is attached to a post, authors often voluntarily retract their posts by deleting them. A 2024 study found that community notes attached to posts on X reduce engagement and diffusion through the social network—although this was later contradicted.

Any potential benefits need to be put in the context of the proportion of notes that get published during the relevant timeframe, and the quality and relevance of those notes within a harm framework.

Risks

X has published a blog about ongoing challenges relating to the Community Notes system, and the steps it is taking to mitigate these, such as preventing coordinated manipulation attempts and low quality contributions putting pressure on ratings. Since introducing AI contributors to the system, it has not updated this page. Neither has X addressed concerns around speed, the lack of a harm framework, lack of coverage of the most harmful content, or its dependence on independent fact checking. Meta needs to respond to these concerns publicly before making changes to Community Notes and TPFC.

1. The need for a harms framework

The potential for a claim to cause harm is the most important commissioning criterion for fact-checkers, followed by concerns like checkability and prominence of claimants. Someone choosing to not spend a dollar with a company, not vote, or not take life-saving medication are examples of actions someone might take based on seeing misinformation. Community Notes cannot engage with the urgency and nuance of this work.

On any given day, there are more than 200,000 potential claims Full Fact could check. Our AI tools narrow this down to around 30, and then we apply a predictive model to understand which claims have the most potential to cause harm. This includes whether a claim could cause a substantively false factual understanding; whether the claim would be seen as true by sufficient people to cause the type of consequence that could be caused if they act on that understanding; and whether those who see the claim as true have the capacity and motivation to act on this false understanding, now or in future, if the claim is repeated.

Previously, Community Notes on X prompted volunteers to classify tweets as being 'believable by many' or 'believable by few' and as posing 'considerable harm' or 'little harm.' However, this feature was removed in 2024. The system has also been criticised for its reductive focus on true-false binaries, which ignores nuanced misinformation and context-dependent harm.

2. Lack of coverage of harmful misinformation

Another significant challenge is that Community Notes frequently leave harmful misinformation uncontextualised. Community Notes' visibility is crucial to their effectiveness. Meta's CISO, Guy Rosen, revealed in September 2025 that just 15,000 notes had been written, and of these only 900 published.

Looking at the data on X raises further questions about the constrained impact of community notes-style programmes. Maldita.es' study during the last EU election concluded that less than 15% of the tweets containing electoral disinformation had a visible community note and, among the 20 most viral debunked posts that received no action from the major digital platforms, 18 were on X with over 1.5 million views each. Even when notes are written, a high percentage are never displayed to users. Maldita.es found, that for any community note that is proposed globally, only 8.3% become visible.

In Mahadevan/Mantzarlis' research on Community Notes on X during the last US election, only 29% of fact-checkable tweets within their sample carried a helpful note. There were many checkable claims that didn't receive any notes, some of which racked up millions of views and appeared on X's election integrity hub. They also looked at Notes created during the three days prior to the election, and found that fewer than 6% of the roughly 15,000 notes reached helpful status, and that only 13% of all notes during this crucial period were even about the election (for example top rated public notes on the theme of whether French has a word for toes, or Bill Clinton's relationship with Monica Lewinsky). Elsewhere, many of the notes were not checkworthy, covering topics like whether Kamala Harris faked calling a voter or a dog picture posted by JD Vance.

Demos has researched the effectiveness of Community Notes at containing the harmful false rumours that directly fuelled the UK's 2024 Southport riots. Demos found that Community Notes were largely invisible to users during the riots, so could not prevent false and harmful information spreading. Only 4.6% of posts in the dataset had Notes created during the Southport riots that were publicly visible during the same period. 78.9% of posts had no visible Community Note, despite 424 having been created during the riots.

The Center for Countering Digital Hate's investigation into how Community Notes operated during the US elections found that "74% of accurate community notes on US election misinformation never get shown to users." This has led critics to challenge whether it is even desirable or worthy to prioritise 'consensus' in moderation systems designed to address misinformation.

3. Slow publication times

Misinformation often spreads much faster than corrections, achieving reach and impact within hours. In contrast, the lifecycle of Community Notes can take much longer, with studies quoting median response times at 18+ hours and 65.7 hours after the original post. A 2023 study found evidence to suggest that Community Notes might be too slow to effectively reduce engagement with misinformation in the early, most viral stage of diffusion. By the time a Note appears, the damage is done.

Demos' research into the misinformation surrounding the Southport riots found that Community Notes were too slow to prevent false and harmful information going viral. The average time between a post being created and a Note being published was 7.8 hours, rising to 19.8 hours on 30th July (the day the riots began). To date, harmful and inaccurate posts created over the period of the riots without a visible Community Note have been viewed 67.6 million times.

4. Reliance on fact checking

Research has repeatedly demonstrated that Community Notes are reliant on professional fact checking. Fundación Maldita.es found that fact checkers are the 3rd most cited reference globally in proposed

community notes on X, and links to fact checkers are present in 1 of every 27 notes proposed. Notes which include fact checking links are also more likely to be published, and faster.

Meanwhile, platforms are publicly moving away from direct financial support and collaboration with these organisations. Without fact checking the system gets jammed. Community Notes cannot replace fact checking as it stands because it is so deeply reliant on it for efficacy, credibility, and timeliness.

2. The suitability and adaptability of consensus or bridging-based algorithms, which are employed in systems like community notes to identify and promote content that appeals across divided audiences, to different political contexts and information environments.

Bridging based algorithms not suitable as a sole moderation system

Bridging algorithms have been used successfully in a deliberative political debate context, but present challenges when within a moderation context. While the theory is attractive, the practice does not match up. Aside from the issues of speed and low publication rates, the bridging algorithm presents other challenges:

- **Ideological gridlock where bridges can't be built**: The bridging algorithm frequently holds notes in limbo, especially when they relate to controversial or politically charged topics. CCDH reported that "74% of accurate community notes on US election misinformation never get shown to users."
- **Information emergencies:** during disasters, attacks, bridging based algorithms are not suitable to contain misinformation that drives harmful behaviour or endangers communities.
- **Dependency on crowd composition:** a recent large-scale analysis showed that the top 10% of contributors produced 58% of notes. If participants do not reflect diverse ideologies and demographics, the arc of the bridge moves accordingly.
- **Potential for AI contributors to clog rating system**: Indicator recently found that an AI bot is now the top contributor to Community Notes on X, and Alexios Mantzarlis' has found that getting a first note published has a positive impact on an author writing further notes. AI Contributors can write but not rate notes, meaning more humans will have to spend time rating notes. The effect of this on clogging up the rating system is as yet unknown, and moreover does not equate to 'crowdsourced' fact checking.
- **Vulnerability to manipulation**: a small minority (5% to 20%) of bad raters can strategically suppress targeted helpful notes, effectively censoring reliable information and raising concerns about reliability.

4. Challenges and best practices in risk assessment, monitoring, and mitigation for the global rollout of social media products, particularly in contexts of polarization, conflict or limited human rights protections.

Full Fact's 2020 report looking at the challenges of online fact checking, including working with global technology companies and using their products, found several challenges which are unfortunately still pertinent and relevant to the global rollout of products today.

Foremost is a lack of transparency and data sharing—an ongoing issue. Recently, a group of VLOPs were approached by Mozilla, the EFCSN and other organisations with a request to share the top 1,000 most-viewed public posts per EU country (made under DSA article 40(12)). This was refused: unsurprising given VLOPs' historical unwillingness to provide data when asked. For example, in 2023, the European Fact-Checking Standards Network Governance Body approached Meta with a request for specific metrics about the impact of the TPFC (for example, the number of distinct fact check articles from EU-based TPFC partners at member state level), and reasons for wanting these metrics. These requests were never granted.

Other VLOPs have also been unforthcoming with information about the impact of fact checks in products, and how/if our work is being used to train LLMs, and how LLMs have been prompted to treat fact checks in generated answers in search and chatbots. In our own work at Full Fact we routinely ask all major LLMs about the claims we are checking before and after we publish them. Near universally, they are used by these models within hours, at times linked and credited and at times not. This culture of hoping AI can be used to scale our way out of integrity problems—propped up by a vanishingly small collection of human-created content—is not sustainable. X's model of using AI to review AI-generated notes is a faded facsimile of a viable model. Important choices need to be made by human experts, and they need to be continued to be funded and celebrated, rather than exploited and ignored.

Our 2020 report also identified challenges around platforms not acting upon feedback about significant problems spotted with products being rolled out, including conversations about CrowdTangle, automated claim matching and content labels . For example, Full Fact has raised concerns about policies based on national borders which do not make sense in a globalised information ecosystem, without success. This is indicative of the power and culture of VLOPs in general. Full Fact recently provided Google with multiple rounds of evidence about misinformation and hallucinations in Google AI Overviews (billed as an 'experimental' product), and we have yet to hear what if any steps the company is taking to mitigate these issues, despite Full Fact providing detailed suggestions and proposals.

5. Research into the efficacy of responses to misleading information beyond content removal, such as fact-checking, labelling, reduced distribution, increased friction, and user-generated context. Additionally, research on avoiding bias in such responses.

While there are legitimate concerns about habituation to or oversaturation with prompts and labels and the decay of interventions' effects over time, there is a wealth of evidence from 2015 onward showing that responses aside from removal are effective in reducing belief in and the spread of misinformation, do not have a backfire effect, and do not polarise audiences. Effective interventions include:

- **Fact-checking**, which must be appraised in light of the organisation's aims (e.g. holding politicians to account, or improving media literacy among the public). Evidence across different continents shows that fact checking reduces false beliefs, and that fact checking scores highest as a trusted source in platform labelling interventions. Meta revealed that 46% of attempted reshares of fact-checked content were abandoned by users, and that 95% of users do not click to view content labelled false.
- **Accuracy prompts**, which have been tested and found effective in the US and beyond.
- **Friction interventions** introduced by platforms such as WhatsApp's restrictions on frequently-forwarded messages in 2020, and read-before-sharing prompts introduced by Twitter in 2020, and Meta in 2021, were found to be effective in delaying the spread of misinformation.
- **Prebunking**, found to be effective in improving audiences confidence in identifying and calling out misinformation and identifying manipulative techniques. It has been championed by Google, but some studies have questioned the efficacy of inoculation outside of a simulated social media environment.
- **Labelling**, which reduces belief in and the spread of misinformation (acting as a social cue even when people do not trust the source), although there is evidence to support the idea of an implied truth effect, where users see items labelled false and assume that unlabelled items have been verified as accurate.
- **Algorithmic demotion**. Demotion is useful to contain misinformation from new audiences, but does not necessarily positively influence the behaviour of offending groups posting misinformation. Recently, academics have argued for content-neutral demotion, focusing on behavioral signals such as posting frequency that exceeds human capability or sharing content that is consistently blocked by other users. This type of system could be less vulnerable to accusations of political bias than some other options outlined above, regardless of their effectiveness.