Public comment by the European Fact-Checking Standards Network on community notes

Community notes have emerged as an additional approach to addressing the spread of misleading or false content. To date, X remains the only platform to have implemented a large-scale, global Community Notes program, and much of the academic and policy-oriented research on the subject draws from its data and experience.

While Meta has announced its own version of Community Notes, it is very limited in scope, though conceptually very similar. According to data shared by Meta's CISO, Guy Rosen, as of September 2025, only 15,000 notes had been written, and just 900 of those were published. This suggests a **considerably constrained program**, raising questions about its reach, representativeness, and overall impact.

As we have previously pointed out to the Oversight Board, misinformation poses significant risks for several fundamental rights. Thus, mitigating misinformation while respecting freedom of speech is essential. (For more details see EFCSN Comment Meta Oversight Board case 2025-050-FB-UA.) **We are concerned that Meta's questions to the board do not reflect this obvious truth. In our view, if Community Notes fail to protect fundamental rights they simply shouldn't be implemented in more countries.** Meta only mentions factors it might consider when deciding which countries to omit from community notes. The company even goes so far as to explicitly ask not to assess the system holistically, thus preempting a necessary and relevant discussion about the efficacy and risks of a community notes system. It is noteworthy that Meta in its policy advisory opinion request to the OB does not disclose any usage numbers which raises questions about the efficacy of the program.

This comment reviews the available research on Community Notes, assessing its efficacy, as we believe this to be the most important aspect. We also compare Community Notes to other interventions such as professional fact-checking, content labelling, reduced algorithmic distribution, increased friction (e.g., prompts and warnings), and user-generated contextual information. Finally, we propose ways to integrate community notes with existing interventions.

Efficacy and risks of community notes

Considering available academic literature, investigations into and reports about community notes, we must conclude that the approach **ill-suited for countering misinformation and disinformation** when it is not complemented by other measures, for the following reasons:

1. **Only a small fraction of submitted notes are ever shown publicly**, even when they relate to posts identified by fact-checkers as containing false or misleading information. The Community Notes Dashboard, developed by CheckFirst, provides a detailed, language-specific overview of the proportion of Community Notes on X that are displayed. The aforementioned data disclosed by Guy Rosen reaffirm that this is also the case for Meta's platforms. Low publication rate might also decrease the engagement of contributors leading to even further reduced impact.
2. **The consensus-based design of Community Notes is ill-suited to polarizing or politicized topics** where consensus is hard to achieve, regardless of factual accuracy. In November 2024, Science Feedback found that most of the content on X that fact-checkers found to be false or misleading had no visible signs of having been moderated. An analysis by Fundación Maldita.es revealed that Community Notes on X was underutilized following the

Valencia floods. Only 8.5% of tweets containing debunked falsehoods had visible accompanying notes.

3. **Community Notes can be manipulated** as this Wired investigation anecdotally points to. In some instances, Community Notes also contained unproven claims. Higher risks are also associated with smaller numbers of contributors: Razuvayevskaya et al. found that the top 10% of contributors produced 58% of all notes. Based on a simulation, Truong et al. concluded that "a small minority (5–20%) of bad raters can strategically suppress targeted helpful notes". Small language areas in which there is only a limited number of contributors, might therefore be at an even greater risk.
4. Related to the previous point, it is an open question **how many users will actually be willing to participate** in contributing and rating notes. According to research by Arjmandi-Lari, Mantzarlis & Stafford the number of monthly active authors on X's Community Notes (defined as users who contributed at least one note a month) has dropped significantly in 2025. The low publication rate of Community Notes might be an important reason for this trend.
5. **Community Notes are too slow** to curb the spread of misinformation. Razuvayevskaya et al. found that "notes, on average, are published 65.7 hours after the original post, with longer delays significantly reducing the likelihood of consensus".
6. Meta has previously announced that **users will not be able to apply Community Notes to advertisements**. This exclusion creates a **powerful incentive for disinformation actors** to simply promote their content as ads—ensuring broader reach and immunity from public correction—while simultaneously generating revenue for Meta. Additionally, it shields scammers and fraudsters from this intervention. Given the enormous scale at which ads promoting scams and frauds are being proliferated, this policy choice should be reversed.
7. Meta has also stated that posts tagged with Community Notes will not be downranked in feeds. This creates a serious loophole: bad actors can repeatedly spread falsehoods, and even if their posts are eventually annotated by users, they **may continue to receive algorithmic amplification and reach**. To our knowledge, being labelled with community notes also does not impede on an account's or page's eligibility for Meta's monetisation programs. This weakens any deterrent against coordinated or habitual disinformation.

Comparison of interventions

The abovementioned shortcomings of community notes for mitigating the spread and impact of misleading information and/or misinformation must be taken into account when comparing community notes to other approaches, most notably to the existing fact-checking program on Meta's platforms. The fact-checking program does not lead to content removals, but instead informs labelling, introduces some light levels of increased friction (at the point of sharing fact-checked misinformation), and may reduce algorithmic amplification of demonstrably false claims: it is both freedom of expression friendly and more effective in curbing misinformation.

- 47% (Facebook) and 46% (Instagram) of **attempted reshares of fact-checked content were abandoned** by users, thereby reducing the spread of misinformation without limiting freedom of expression (see Meta's Code of Practice Transparency Report, March 2025, SLI 31.1.2).
- **95% of users do not click to view content labeled false** (source), indicating that users do not want to be exposed to false claims.
- Less than 4% of appealed fact-checks result in reinstatement on Facebook, compared to over 60% for other moderated content, indicating **high accuracy and reliability** (see Meta

DSA Transparency Reports accessed through https://transparency.meta.com/reports/regulatory-transparency-reports).

Independent research supports the effectiveness of warning labels, showing significant reductions in belief and sharing of false content, even among users skeptical of fact-checkers (for an overview see Martel & Rand, 2023).

A common critique levied against professional fact-checking is the allegation of political bias, often advanced by political actors who claim that fact-checkers disproportionately target certain segments of the political spectrum. These claims, however, are rarely substantiated by robust empirical evidence. On the contrary, research has demonstrated how politically-balanced groups of average users consistently reach the same conclusions about a claim as professional fact-checking organisations. Independent fact-checkers, particularly those adhering to industry frameworks such as the European Code of Standards, operate under publicly available commitments to transparency, non-partisanship, and methodological rigour.

In contrast, community-based systems such as X's Community Notes do not rely on institutional oversight to safeguard impartiality, instead using "bridging" algorithms. While this approach is often framed as a safeguard against bias, empirical studies suggest that **Community Notes may themselves reflect skewed inputs**. For example, one study found that community notes on X contained a disproportionate number of left-leaning sources. Another study highlighted asymmetries in which political actors of one side are more frequently flagged. Importantly, such findings should not be interpreted through the lens of false equivalence or "bothsideism." Research has repeatedly shown that the distribution of false or misleading information is not necessarily symmetrical across the political spectrum. Therefore, any evaluation of bias in content moderation or contextualisation systems should prioritise factual accuracy and methodological transparency over perceived political balance.

Integrating community notes with existing interventions

Considering the prevalence of misinformation online and its detrimental effects on fundamental rights, **it is appropriate for digital platforms to run both a fact-checking program and a community notes program simultaneously**. Such programs could either be independent of one another or they could be integrated. Fact-checking can fruitfully complement community based approaches by inputting professional, independent skills and expertise in the information mix. In fact, many community notes reference fact-checks to establish helpful context or rebuke falsehoods. Furthermore, the design of community notes can inform the design of fact-check labels. Drolsbach, Solovev & Pröllochs showed that the context and explanations provided in community notes improved the trustworthiness of notes significantly compared to more limited flags.

Fact-checkers can contribute to crowd-based notes system in various ways, keeping in mind that time and accuracy are both of the essence: by suggesting verified, high quality notes themselves; by checking the crowds' notes to accelerate their visibility to other users; by offering innovative, technical solutions to quickly match claims and verified notes. We would encourage Meta to contemplate the integration of its professional fact-checking into community notes. **Extending the existing Third-Party Fact-Checking Program** to also contribute to Community Notes is a promising way to improve information integrity and safeguarding freedom of expression and information on Meta's platforms. Specifically, we recommend the following:

- **A "fast lane" for fact-checkers**: A consensus voting system may suppress valid notes if they benefit one side of a partisan debate. A fast lane for certified fact-checkers could take

different forms, e.g. auto-approving notes by fact-checkers, weighing their votes more heavily, or ensuring fact-checkers crosscheck each other's notes.

- **Quick access for better results**: Giving fact-checkers access to all the proposed or requested notes in real time enables comprehensive input. User notes can be an early warning sign and provide useful clues. But professional fact-checkers are better equipped, after years of experience, at collating, researching, evaluating and summarizing the available evidence into a clear conclusion.
- **Transparency on partnerships**: Notes could be utilized to make the partnerships between platforms and independent fact-checkers more transparent to users.
- **Avoid bias**: Professional IFCN or EFCSN certified fact-checkers adhere to standards on non-partisanship and independence. They must prove that they are compliant with these principles and this compliance gets reevaluated on a regular basis.
- **Independence is key**: Ensuring that fact-checkers are fairly remunerated for the work provided while remaining editorially independent is a way to ensure sustainable, reliable quality of the online service on information access.

In conclusion, although community notes seem conceptually promising as a crowd-sourced solution to the perceived limitations of other forms of moderations, it remains too limited, and easily manipulated to serve as an effective standalone solution for combating misinformation. Evidence from both X and Meta shows extremely low publication rates, slow turnaround times, vulnerability to strategic misuse, and weak coverage of polarizing topics, all of which diminish the impact of community notes. Meta's current implementation raises additional concerns: the program is small in scale, excludes advertising content, and lacks transparency around usage. These shortcomings create loopholes that are easy for bad actors to exploit.

Professional fact-checking on Meta's platforms has demonstrated clear effectiveness in reduction of circulating misinformation through accurate labeling, downranking and reduced resharing, confirmed by user behaviour data. Our recommendation is to integrate community notes with the established Third-Party Fact-Checking program, for example by providing fact-checkers with a "fast lane," visibility into user-submitted notes, and mechanisms to accelerate high-quality contextual information. By highlighting the independence of fact-checkers and transparency of the community notes system, a blended approach leverages the strengths of both crowdsourced insight and professional expertise to reduce misinformation while protecting information integrity and the fundamental right to freedom of expression.