December 8, 2025

To the Oversight Board,

We submit this comment as researchers at Georgetown University and Cornell Tech who study adversarial abuse, content moderation, and platform governance. We welcome the Board's call for comment on Meta's request regarding the international expansion of Community Notes. But the way Meta has framed the request is, in practical terms, unanswerable because the company has failed to publish any meaningful data about the program's "pilot" phase in the United States. Extrapolating findings from X's Community Notes can only take us so far given the way data from that program is structured and the significant difference between X and Meta's platforms. The Oversight Board should decline to offer an advisory opinion until significantly more information is transparently provided – both to the Board and to the public.

In its request, Meta asks the Board to advise on factors to inform which countries should be omitted from the global roll-out of Community Notes, while explicitly instructing the Board *not* to consider "general product design or the operation of the Community Notes algorithm." However, those are precisely the things that determine whether the system is safe, effective, and compatible with Meta's human-rights responsibilities. A country-selection framework that does not consider how the product actually works is a framework for rubber-stamping, not for oversight.

The Board should decline to accept those constraints. If Meta wants the legitimacy that comes from an independent policy advisory opinion, it must accept scrutiny of the product itself and provide the evidence necessary for that scrutiny.

1. The question Meta has posed is structurally flawed

Meta's request asks the Board to identify "factors we should consider when deciding which countries, if any, to omit from the international roll out," and "how to weigh those factors," while *also* asking the Board not to opine on product design or how the algorithm operates.

However, whether Community Notes *reduces* or *amplifies* harm in a given context depends on the interaction of:

- The design of the rating and publication algorithm;
- Who participates (and who is excluded) as contributors;
- How notes are interpreted by users;
- How the feature interacts with Meta's other recommendation and enforcement systems.

Those are design questions. They cannot be bracketed off as irrelevant to a "country-level" analysis. For example, if there are systematic coverage gaps (no notes on content in minority languages, or very low uptake in closed Groups), that matters far more in some countries than others—but outsiders cannot reason about that without any transparency.

The Board has previously insisted, in its Cross-Check policy advisory opinion, that system-wide design choices *are* part of its remit. There is no principled reason to treat Community Notes differently.

Under the EU Digital Services Act (DSA), large platforms must assess *systemic risks* arising from "design, functioning and use" of their service—including negative effects on civic discourse and electoral processes— and adopt proportionate mitigation measures. If Meta intends to cite the Board's opinion in its engagement with regulators or civil society, the Board should be explicit that no such separation along design or usage lines is possible here either.

Put bluntly: the Board cannot give a meaningful answer to "where" without looking at "what" and "how."

2. Meta's pivot away from fact-checking is political, not merely technical

On January 7, 2025, Meta announced that it would end the third-party fact-checking program in the United States and "move to a Community Notes model." In that same blog post, Mark Zuckerberg framed the change as a response to "experts" with "biases and perspectives," arguing that fact-checking had become "a tool to censor" legitimate political speech.

This was explicitly political messaging, not a neutral audit of program performance. No evidence from eight years of running the fact-checking program was provided by Zuckerberg to support this claim. Researchers have found that politically asymmetric interventions against misinformation can result from politically asymmetric sharing of misinformation. They've also found that Community Notes *also* displays a skewed distribution of fact-checks by political background, suggesting something bigger than "fact-checker bias" may be at play. Meta's statements that its fact-checking program was "biased" are contested political claims, issued in a charged political context. They should not be treated by the Board as evidence-based assertions that justify a wholesale pivot to a crowdsourced model.

That context matters for the Policy Advisory Opinion Request, given that Meta is asking the Board to bless the next phase of that pivot—global expansion of Community Notes—without offering any meaningful transparency into what will replace it in practice.

3. What we actually know from X's Community Notes—and why transparency matters

Meta repeatedly cites X's Community Notes as its model, saying at launch that, "We won't be reinventing the wheel. Initially we will use X's open source algorithm as the basis of our rating system." In its request to the board, Meta explains it has altered X's rating algorithm somewhat to separate "helpfulness" and "consensus" scores. This suggests that consensus matters *less* on Meta's Community Notes than on X, but the company only shared a very basic explanation of its algorithm and pointed to a research paper by X for more details.

X's program has been operational for nearly five years and was built on a foundation of ***unusual transparency and auditability*****,** which Meta has not replicated. From its launch as Birdwatch, X's Community Notes team has published a comprehensive, regularly updated dataset of notes and ratings, as well as the open-source code for the core "bridging-based" consensus algorithm. This has made it possible for outside researchers and journalists to evaluate coverage, accuracy, political balance, abuse attempts, and even the "data labor" dynamics of contributors. Meta has done neither.

Based on this transparency and auditability, researchers (including the authors of this comment) have highlighted **mixed but informative empirical results.** Studies and early analyses suggest that:

- Community Notes can reduce engagement and reach of labeled false or misleading content, sometimes substantially, once a note is attached and seen.
- The "people who usually disagree" requirement helps reduce obvious partisan slant, but also means that controversial content is *less* likely to get a note at all.
- The system is potentially vulnerable to coordinated manipulation, including organized campaigns to mass-down-rate accurate notes or to stack contributor pools in particular languages or regions; X has acknowledged this concern publicly.
- Note authors are not professional fact-checkers and are required to share links to support their claims. Those links often point to media organizations, Wikipedia, and fact-checkers. Rolling out Community Notes on Meta *in lieu of* fact-checking partnerships may make the former less effective by starving contributors of relevant links. Here, too, having data from Meta on what sources are being used by "helpful" note writers would help provide better guidance.
- Meta asks for country-level information about Community Notes that it won't share and that X's data is ill-suited to extrapolate from. X's Community Notes contributors aren't tagged by location, so it is necessary to rely on proxies like language to try and assess this request. This only works for languages that are significantly tied to one country alone. A rudimentary analysis conducted by one of the authors of this comment found that indeed there appear to be divergent geographic dynamics in consensus-building and contribution rates within X on Community Notes. Specifically, Japanese-language users appear to

engage more than the average and get notes rated helpful almost 40% more often than the global average. We can speculate as to what is going on, but more research is necessary and none is available with the rigor and detail necessary for Meta to make a determination of which countries to include and exclude from its own program.
- One preprint found that in countries with more multipolar polities, the Community Notes bridging algorithm may result in slightly higher rates of "Helpful" notes as cross-over agreement on specific topics is more possible than in strict bipolar societies like the United States. This may be something to consider when expanding and testing Meta's Community Notes program, but it is based on data *from another platform*.

The material that Meta submitted to the Oversight Board about its pilot program, which has been operational for approximately 9 months, contained no statistics or information about uptake. There is no public, regularly updated dataset of notes, ratings, or contributor behavior that would allow independent evaluation that might inform an understanding of what to do next.

Early reporting on Meta's version of Community Notes by one of the authors of this comment highlights ongoing uncertainties about who gets access, how often notes appear, how they are ranked and presented in feeds, and how abuse is handled—precisely the kinds of questions X's transparency has allowed researchers to examine. This is not an insurmountable problem, clearly. But it highlights the problem with this request: without X-level openness, neither the Board nor outside experts can adequately assess what facets of Meta's implementation will impact unique populations globally.

4. Unpaid "data labor" and the need for fair compensation

Our final point speaks to incentives: Meta has publicly emphasized, including in filings to the European Union, that it has spent over $100 million supporting professional fact-checking partners since 2016. Ending the U.S. program and shifting to Community Notes does not make that work go away; it simply reassigns it to volunteers.

Research on Birdwatch/Community Notes has framed contributors' work as "data labor"—a form of unpaid annotation that directly improves the platform's products and mitigates harms experienced by other users. At the same time, Meta is expanding its investments in creator monetization and AI-generated content, encouraging users to produce highly engaging material, some of which may itself require correction or contextualization.

If Meta is serious about building a durable Community Notes ecosystem, the Board should:

- Encourage Meta to adopt clear, transparent policies for compensating or otherwise materially supporting contributors who perform substantial moderation and contextualization work.

- Ask how Meta will ensure that this unpaid labor is not disproportionately extracted from certain communities (e.g., journalists, local experts, or marginalized groups) without corresponding investment in sustainability.

This issue goes to the integrity and resilience of the system Meta is asking the Board to endorse.

Summary of Recommendations to the Board

Given the above, we urge the Board to:

1. **Reject Meta's attempt to exclude product design and algorithm operation from the scope of review.** The Board should state clearly that it cannot provide meaningful guidance on country-level roll-out without assessing the design, performance, and governance of the preliminary Community Notes program deployed in the United States.
2. **Tie any recommendations to concrete transparency commitments.** At minimum, before expanding to additional countries, Meta should:
    - Publish a regularly updated dataset of Community Notes and ratings (with appropriate privacy protections), similar to X's releases, and
    - Open-source the core ranking/consensus logic, or provide functionally equivalent documentation
3. **Explicitly decline to endorse Community Notes as a shield for the DSA and other regulatory processes until more is known about the program's efficacy.** The Board should make clear that its advice is not an endorsement of Community Notes as a sufficient systemic risk mitigation under the DSA or similar laws, and that regulators should continue to demand compliance under any existing arrangements.

Nine months after the program's launch, Meta has not yet been transparent with the public about how Community Notes actually functions across its properties, or how it will be evaluated. It made a political decision, and now asks the Oversight Board to validate the next step of a significant pivot, while explicitly cordoning off scrutiny of a new and untested system's design.

The Board should refuse that framing. Instead, it should use this PAO Request as an opportunity to reaffirm that meaningful oversight of platform governance requires commensurate transparency.

Thank you for the opportunity to submit a comment on this important topic.

Renée DiResta and Alexios Mantzarlis