Community Notes
Policy Advisory Opinion
PAO-2025-01

Executive Summary

The Oversight Board finds that community notes could enhance users' freedom of expression and improve online discourse on Meta's platforms if implemented with sufficient scale, speed and safeguards against manipulations. However, in certain circumstances – including in repressive human rights regimes, in particular electoral contexts and in ongoing crisis and conflict situations – expanding community notes to countries outside the United States could also pose significant human rights risks and contribute to tangible harms that Meta has a responsibility to avoid or remedy.

The Board is also concerned about coordinated disinformation networks potentially abusing community notes, and the risk that community notes could, in certain contexts, privilege dominant political, ethnic or linguistic groups, and marginalize minority groups.

The likelihood and severity of these potential human rights risks and their adequate mitigation depend greatly on the design and functionality of the community notes product in each context. The effectiveness and adequacy of Meta's mitigation measures regarding some of these risks – for example, to ensure contributor anonymity and protect against those who try to game the system – will need to be verified through an ongoing process of data gathering and reporting on how community notes function in practice.

In addition, insofar as Meta envisions community notes as its primary way to address misinformation falling short of its threshold for removal (i.e., where there is not a likelihood of contributing to the risk of imminent physical harm or to interference with the functioning of political processes), the Board finds that the program's design may limit its ability to accomplish that goal. Delays in note publication, the limited number of published notes and its dependence on the broader information environment's reliability raise serious doubts about the extent to which community notes can meaningfully address misinformation linked to harm.

The Board recommends criteria that Meta should use to assess when these human rights risks may warrant withholding community notes from a particular market. The Board's recommendations are necessarily conditional because Meta's questions cannot be addressed conclusively without sufficient testing and detailed data about how the community notes algorithm functions in real-world situations and in relation to other misinformation tools. For that reason, the Board also recommends ongoing data gathering, assessment and reporting regarding the functionality of community notes, related to those criteria.

Background

On November 19, 2025, the Board announced that it had accepted a request from Meta for guidance on the specific factors the company should consider when deciding whether any country should be omitted from its planned expansion of community notes outside the U.S., as local context might impact the program's operations. Additionally, Meta asked the Board how to weigh such factors in relation to one another, in a way that can be applied on a large scale.

In its request, Meta said the community notes program is in an "early stage of product development" and the company possesses "limited data from the U.S. beta rollout." Meta described this rollout as a period of "testing and refinement" that may result in the form of community notes evolving. Because of these considerations, the company's primary interest lies in establishing "fundamental guiding principles" for its implementation worldwide. Importantly, the Board has not evaluated the general effectiveness of community notes in the U.S.

The Board consulted a range of stakeholders, including technical experts on bridging algorithms, civil society organizations, journalists and fact-checkers, for observations and data on different community notes-style moderation systems (including X's) across different contexts.

On January 7, 2025, Meta announced it was introducing a community notes program and ending its third-party fact-checking program in the U.S. The company indicated it would refine community notes before making it available for users outside the U.S.

Key Findings and Recommendations

Informed by a broad consideration of potential human rights impacts, the Board has developed the following country-level factors to guide Meta's rollout of community notes:

Safeguards in repressive human rights contexts

Community notes depends on an active and engaged contributor base. Such programs have experienced greatest success amid a robust civil society. Ideally, contributors rely on independent media to support proposed notes and participate without fear of harassment or retribution.

The Board underscores that until Meta can demonstrate robust and effective contributor privacy protections, with evidence of red-teaming under adversarial conditions, a clear policy on handling requests from law enforcement agencies for community notes data and risk mitigation measures, countries with repressive human rights records and weak civil societies should be omitted from the initial rollout.

Exercise caution during elections

Community notes can support access to information and freedom of expression during elections in robust information environments with free media and uninhibited civil society. Without those conditions, the program risks publishing misleading notes, and Meta should proceed with caution. Where significant risks to the integrity of political institutions are present, and Meta determines through product testing, risk assessment and human rights due diligence that its safeguards are insufficient to mitigate them, community notes should not be introduced in advance of or during major elections.

Omit countries with a history of coordinated disinformation networks

Community notes operates under the assumption that a sufficiently diverse and independent set of contributors will evaluate content in good faith and that consensus signals can reliably approximate accuracy. Where malicious actors have repeatedly demonstrated the ability to coordinate large numbers of accounts to promote deceptive information, this assumption may not hold. Instead, community notes risks becoming a vector for manipulation rather than a safeguard against it. This risk will become more acute as artificial intelligence facilitates the scaled creation and operation of these networks.

Until such time as data on the functionality of community notes can verify the adequacy of Meta's mitigation measures against such coordinated activity, the Board recommends that Meta initially omit countries with a historical pattern of intentional, large-scale disinformation networks. Meta should also consider here whether actors have demonstrated the intent to manipulate information ecosystems and possess the technical sophistication to do so on a large scale.

Do not introduce in crisis or protracted conflict conditions

The potential vulnerability of community notes to coordinated manipulation by armed groups, state actors or their supporters seeking to legitimize propaganda through gaming of the note rating system poses heightened risks in ongoing crisis or conflict situations. Conditions causing groups' inability to contribute, e.g., unstable internet access and insecurity preventing participation, can exacerbate information asymmetries. In such circumstances, timeliness is critical. Delays in note publication suggest community notes may be an inadequate primary safeguard in crises and conflicts, particularly as thresholds for incitement to violence may be lower. Notes targeting specific groups can more easily result in offline harm.

Due to uncertainty over community notes' performance in conflicts and its potential to heighten risk of harm, the Board believes that it should not be introduced in countries experiencing crises or protracted conflict.

Delay introducing community notes where there is language complexity that Meta cannot technically and operationally accommodate

Community notes requires sufficient representation of a context's language groups. Where Meta is unable to achieve this, it should delay introducing community notes. Otherwise, language disparities in notes proposed and published could be created or exacerbated, undermining the program as a plural and diverse information source.

Additionally, there are potential linguistic and cultural variations in the use and interpretation of community notes, some unexpected, which may affect which notes are published. For example, what it means to rate something as "helpful" might differ in different places and when the word is translated. In countries where Meta anticipates features of community notes cannot yet accommodate linguistic complexity, it should consider delaying its introduction.

Exercise extreme caution where social division and disagreement that drives political violence cannot be outlined simply

The design of X's community notes algorithm implicitly assumes disagreement and division in a particular context can be modelled simply, giving a single measurement of polarization. Meta has not provided any information that suggests its program will be different. However, where division and disagreement cannot be easily modelled along a single axis, this can reinforce an unduly simplistic understanding of conflict, overlooking how multiple factors intersect (across politics, ethnicity, religion, language and caste, for example). In practice, this risks marginalizing minority perspectives, as misleading or harmful notes reaching consensus among majority groups may be published. Where the algorithm does not accurately identify and attempt to "bridge" (rewarding content receiving positive feedback from audiences that typically disagree) a division that drives conflict and violence, this risk of harm is especially acute. Therefore, the Board recommends that Meta exercise extreme caution when considering countries characterized by these dynamics.

Where there is a heightened risk of misalignment between the design assumptions of community notes about a society and what actually drives social and political division there, such countries need not be categorically omitted from community notes' rollout. Rather, the rollout should be sequenced to allow performance testing in different contexts, and should begin cautiously in political, linguistic, and information environments that are similar to places where Meta already possesses such data, piloting and risk mitigation measures.

Omit countries that face persistent obstacles to internet access

The Board recommends omitting countries that face persistent or systemic obstacles to internet access, as community notes relies on broad, consistent and equitable contributor participation to function as intended. A narrow pool of contributors, due to infrastructure gaps, high costs, regional disparities and, especially, government-imposed shutdowns, undermines the core premise of community notes' representativeness.

Weighting

The Board recommends that factors whereby community notes could create or heighten the risk of harm should be weighted heavier than factors from the program's inadequacy as a harm mitigation measure. These include thresholds whereby Meta should not introduce community notes until it demonstrates it can do so while mitigating these harms.

Meta should provide the Board with the criteria or risk matrix it develops to guide expansion every six months during initial expansion, with evidence of how these are applied in country-level decisions about program expansion.

Data and reporting

The potential human rights risks of and adequacy of mitigation strategies for community notes depend on the design and functionality of the product in context. Meta told the Board it plans to test the program prior to launch in a market. The Board recommends this testing should focus on surfacing and mitigating risks related to contributor anonymity, coordinated disinformation campaigns and gaming of the system, language representation and contributor participation. The Board calls for substantial transparency, reporting and researcher access to data on Meta's community notes performance.

Policy Advisory Opinion in Full

I. Meta's Request

On January 7, 2025, Meta announced that it was introducing a community notes program and ending its third-party fact-checking system in the United States. At the time, Meta indicated that it would refine community notes before making it available for users outside the United States. Meta describes community notes as allowing users to add labels with additional context to content where they believe additional information may be helpful. Participants in the community notes program, called "contributors," rate proposed notes as "helpful" or "not helpful" and explain their response by selecting a reason from a list of options. An algorithm calculates a score that reflects whether a note has been rated helpful among a sufficient number of contributors who usually disagree with each other based on past ratings. If this score reaches a certain threshold, and the note does not violate the Community Standards, then the note will be published. This system contrasts with the third-party fact-checking program, which relies on partner organizations to review, research and make judgments about the veracity of content that can result in labeling false or misleading information.

Meta requested guidance from the Board on the specific factors it should consider when deciding whether any country should be omitted from its planned expansion of community notes outside the United States, as local context might impact the program's operations. Additionally, Meta has asked the Board how to weigh such factors in relation to one another, in a way that can be applied on a large scale.

In its request, Meta said that the community notes program is in an "early stage of product development" and that the company possesses "limited data from the U.S. beta rollout." Meta described this rollout as a period of "testing and refinement" that may result in the form of community notes evolving. Because of these considerations, the company's primary interest lies in establishing "fundamental guiding principles" for its implementation worldwide. The Board has accordingly not undertaken a general evaluation of the effectiveness of community notes as it has operated in the United States.

The Board identified a set of human rights risks that it is most concerned about mitigating in advance of the global rollout of community notes. To do this, the Board consulted a range of stakeholders, including researchers, technical experts, civil society organizations, journalists and fact-checkers, for observations and data on the operation of different community notes-style systems of moderation (including X's) across different contexts. The Board's evaluation focuses on the anticipated performance of Meta's community notes system if applied in other markets beyond the United States in light of that information.

Answering the question of what community notes and the algorithm that underpins it will do in a given context depends significantly on empirical data. Meta indicated that the system will be piloted outside of the United States, with performance data evaluated against Meta's goals for the program. Meta said that it plans to publish downloadable data on community notes, with data categories on par with what X shares for its community notes program. This would include the full text of notes and their associated geographies, ratings for each note, status history, user enrollment and the requests for notes on content. However, that level of data is not yet available for the Board's use, and instead, a set of high-level metrics was provided confidentially to the Board. Meta provided the Board with certain data on community notes, including but not limited to the number of contributors enrolled, the number of notes written and published, and the average time to consensus in the United States. Therefore, this opinion cannot offer empirical analysis of the successes, failures or impact of Meta's community notes program. It does, however, incorporate insights from research on similar crowdsourced programs to understand likely areas of risk in different contexts, which directly relate to the questions Meta posed to the Board.

Importantly, Meta did not ask the Board to evaluate the effectiveness of its third-party fact-checking program, nor the relative merits of community notes and third-party fact-checking. This opinion therefore, does not include independent recommendations about whether third-party fact-checking should be discontinued or maintained. However, given that community notes is included in Meta's approach to misinformation, the opinion does consider the absence or presence of third-party fact-checking as a relevant factor in how Meta addresses deceptive content in a given context. The Board asked Meta if this list of countries where Meta has agreements in place to provide third-party fact-checking will remain accurate during and after any rollout of community notes beyond the United States. In response, Meta said that it is "subject to change periodically." Recent media reporting has found that while Meta has not made a public statement on the future of third-party fact-checking outside of the United States, contracts with fact-checking agencies have been renewed for 2026, albeit at reduced funding levels.

The questions posed in Meta's request indicate that Meta is actively considering extending community notes to countries outside the United States. The Board's opinion neither endorses nor opposes this approach. Rather, its analysis proceeds by analyzing what the implications would be if Meta were to make community notes available everywhere and addresses what considerations should preclude or condition launching the product in a particular country.

The Board accepted Meta's request on November 19, 2025. Meta's full request to the Board can be found here. Following this acceptance, the Board sent Meta questions. The Board asked questions on: how Meta envisions the relationship between community notes and other approaches to deceptive information; how the company defines and measures success for community notes; the risk mitigation measures and human rights due diligence Meta undertook in advance of the community notes launch in the U.S., and similar efforts planned for international launches; the performance of community notes in the U.S.; technical differences between Meta's community notes and X's community notes; and the relationship of community notes to election-related and crisis- and conflict-related measures.

II. Stakeholder Engagement

Over the course of developing this policy advisory opinion, the Board engaged with external stakeholders in the following ways:

Public Comments

The Board received 23 public comments in December 2025 related to this policy advisory opinion that met the terms for submission. Eight were submitted from United States and Canada; five from Asia Pacific and Oceania; five from Europe; three from Latin America and the Caribbean; and two from the Middle East and North Africa.

The issues covered by the submissions included:

- **Potential benefits of community-based moderation to address deceptive content.** Several submissions highlighted the positive impacts of community notes on freedom of expression, such as the adding of context and the system's potential to build information literacy, especially when compared to interventions that result in content being removed (see, for example, PC-315688, Full Fact). These submissions argue that community notes and third-party fact-checking could work as complementary tools. For example, including independent fact-checkers in the program "could be a way to increase the accuracy, the relevance and the streamlining of crowd-sourced notes …

while allowing greater interactions with online users for increased trust" (PC-31579, Agence France-Presse).

- **Contextual risk and uneven global applicability of community notes.** Numerous submissions stressed that political repression, democratic backsliding, conflict and coordinated manipulation fundamentally shape whether community-driven moderation can work safely (see, for example, PC-31580, Dr. Yohannes Eneyew Ayalew and Dr. Maria O'Sullivan). Commenters warned that in countries with a restricted civic space or high surveillance, community notes could be captured by state-aligned actors or malicious groups, distorting consensus and reinforcing official narratives rather than providing opportunities to challenge them (PC-31586, Asian Forum for Human Rights and Development). Several submissions urged Meta to avoid treating global expansion as a simple rollout, emphasizing instead the need for country-specific feasibility studies, human rights impact assessments and minimum thresholds for performance of the system before deployment.
- **Language and linguistic inequality.** Many comments highlighted that community notes requires sufficient contributor density and shared linguistic understanding to generate timely consensus (see, for example, PC-31570, Dr. Sanjana Hattotuwa; PC-31589, Digital Democracy Institute of the Americas; PC-31581, Trusted Information Alliance). Several submissions presented evidence regarding X's community notes, suggesting that notes in non-English languages are far less likely to be rated or published, leaving large user populations underserved. Comments also noted that multilingual posts, code-switching (alternating between languages) and reliance on machine translation introduce risks of misunderstanding and exclusions, particularly in multilingual regions.
- **Limitations in effectiveness and speed.** Empirical studies cited in several submissions suggest that only a small fraction of proposed notes in community notes systems are ultimately published, often after significant delays. As a result, community notes tend to appear on "soft" or low-stakes content, while highly polarizing or fast-moving misinformation, particularly during elections, conflicts and crises, often goes unaddressed (PC-31597, Institute for Strategic Dialogue). For these commenters, this raises concerns that the system may systematically fail where harms are greatest.
- **Relationship between community notes and professional fact-checking.** Across public comments, there was opposition to replacing third-party fact-checking with community notes. Numerous public comments argued instead for a hybrid approach, where community notes complements professional assessments of veracity and media

literacy efforts. Some public comments offered proposals for how this could work in practice. For example, a "fast lane [in community notes] 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 cross-check with each other's notes" (PC-31587, European Fact-Checking Standards Network). Fact-checkers, in particular, warn that withdrawing institutional support, while expanding reliance on contributors, risks weakening the overall information ecosystem.

- **Labor, safety and accountability.** Submissions also raised concerns about harassment, retaliation and personal risk for contributors. One public comment (PC-31578, Renée DiResta and Alexios Mantzarlis) argued that writing and rating notes should be understood as unpaid "data labor" that benefits platforms, and that it risks being extracted disproportionately from journalists, experts and marginalized communities.

Taken together, these themes suggest cautious interest in the community notes' participatory model, coupled with concern about overreliance on the model without stronger safeguards and transparency.

Stakeholder Roundtables

The Board engaged in further stakeholder consultations through two stakeholder roundtables on the opportunities and risks of community notes. These roundtables included approximately 30 participants from the Americas, Africa, Asia and Europe. These included researchers focused on the spread and impact of deceptive information, fact-checkers, technical experts, civil society actors and human rights advocates.

The roundtable discussions were supportive of how the community notes system may give context rather than remove content. In that sense, it can strengthen freedom of expression by shifting moderation from removal to contextualization. Rather than suppressing speech, the system may add counter-speech, allowing expression to remain visible while improving the information environment around it.

Participants also expressed skepticism about community notes as a primary tool for addressing misinformation, particularly in non-U.S. and high-risk contexts. Participants reported that consensus-based systems can fail on polarizing, political or crisis-related content, and may privilege majority power over factual accuracy.

Linguistic diversity and low digital literacy were also flagged as structural barriers, with evidence that non-English and minority-language notes are far less likely to surface, leaving large populations unserved. Stakeholders warned that conflict-affected and authoritarian contexts pose acute risks, including state or coordinated actor capture, harassment of contributors and offline violence, exacerbated by delays to notes being published.

Across the discussions, there was strong agreement that community notes should complement, not replace, professional fact-checking, which can provide the expertise and harm assessment that community notes sometimes lacks (see, also PC-31568, Full Fact). Finally, participants stressed the need for transparency and data access for researchers, urging piloting, independent evaluation and clearer accountability before any global expansion.

III. Meta's Community Notes Program

a. Approaches to Misinformation

Meta describes its current approach to apparently false or misleading information as involving three strategies: removing certain kinds of harmful misinformation; reducing the distribution of content rated false, altered or partly false by third-party fact-checkers; and applying additional information in the form of different labels to content that is potentially misleading or confusing. According to Meta, community notes falls into this third category. While this suggests that one of the primary reasons for community notes is to counter misinformation, answers to the Board's questions indicate that Meta also views community notes as a tool for adding context more broadly. This broader framing reflects a design choice to prioritize contextualization over content removal, allowing speech to remain visible while facilitating users' ability to assess and interpret information.

Meta's Misinformation policy specifies that the company removes false or misleading information "where it is likely to contribute to the risk of imminent physical harm" or "likely to directly contribute to interference with the functioning of political processes," including elections and censuses. To make these determinations around harm, Meta partners with "independent experts who possess knowledge and expertise to assess the truth of the content and whether it is likely to directly contribute to the risk of imminent harm." In its request to the Board, Meta stated that its January 7, 2025, announcements did not affect its continued removal of content that violates these policy lines. This type of content is typically identified and evaluated in consultation with Meta's Trusted Partner Channel (see Haitian Police Station Video decision).

Meta’s external fact-checking partners – in those countries and languages in which they are operating – prioritize assessing potentially false information that is “timely, trending and consequential.” Content rated “false,” “altered” or “partly false” by fact-checking partners may be demoted, not recommended and rejected for ads. Meta’s policy provides that notices are applied to posts with these ratings, and Meta sends notifications to the users who posted them. Content that Meta exempts from fact-checking covers “content that doesn’t include a verifiable claim,” “opinion and speech from politicians” and “digitally created or edited media containing one of Meta’s artificial intelligence transparency labels or watermarks on the basis of its authenticity,” where the misleading aspect is disclosed by the posting user. There is an appeal mechanism to contest fact-checkers' conclusions.

In the United States, community notes contributors can propose notes that add more context on any piece of public, organic (i.e. not paid), content on Meta’s platforms. The public display of the proposed note is contingent on the note surpassing the “helpful consensus” rating threshold and not violating any Community Standards. This includes types of content not previously covered by third-party fact-checking, such as opinions and posts from politicians. Meta states that community notes contributors in the United States (where third-party fact-checking has been discontinued) are part of its approach to misinformation: contributors can “write and submit a note to posts that they think are potentially misleading or confusing.” As explained in more detail below, if contributors who have previously disagreed find the note helpful, it is publicly affixed to the post. The addition of a note does not carry consequences such as demotion or removal from recommendations, as false, partly false or altered fact-check labels do. After the addition of a published community note to a post, Meta also sends notifications to people who have previously reposted, left a comment, rated or requested a note for that post.

In response to a question from the Board about the relationship between community notes and third-party fact-checking, Meta clarified that it views the two as “fundamentally different in several important ways.” For Meta, “third-party fact-checking is designed to address ‘fact-checkable’ claims on its platforms, focusing strictly on statements of fact rather than opinions.” That distinction is crucial for Meta, and points to the wider purpose of community notes: “Community notes contributors are able to add context to any piece of content where they believe additional information may be helpful, including content that would be categorized as opinion or would otherwise not qualify for review under the third-party fact-checking program.”

Another key difference, according to Meta, is that community notes do not carry the “punitive consequences associated with the third-party fact-checking program.” Specifically, there are "no strikes for posting content that receives a community note," as would be the case with posting misinformation that is "likely to contribute to the risk of imminent physical harm," and the "distribution or monetization of such content is not affected."

b. How Community Notes Works

i. Technical Aspects – Algorithm and Queue

In its request, Meta disclosed that it built its community notes system using the open-source algorithm from the community notes program of social media platform X. Meta described the algorithm as a "consensus algorithm that uses separate measures of 'helpfulness' and 'consensus' to calculate an overall 'helpful consensus' score." Consensus algorithms, also known as bridging algorithms, are designed to promote content or interactions that supposedly foster understanding and agreement across divided groups. Theoretically, they reward content that "bridges" – that is, content that appeals to or receives positive feedback from audiences who typically disagree. The consensus-based design of community notes makes it functionally different from third-party fact-checking in its aims and its comparative effectiveness for different tasks.

Meta explained that the algorithm calculates this score by identifying agreement that a note is 'helpful' among a certain number of contributors who usually disagree with each other based on past ratings. According to Meta, if the combined "helpful consensus" score on a note exceeds a "certain threshold" and the note does not violate Meta's Community Standards, the note will be published. In its request, Meta said that because its algorithm is "more complex" in practice, it is "not possible to provide an exact number of how many helpful ratings a note must receive or how much disagreement is necessary for the algorithm to determine contributors have differing viewpoints." The public code from X's community notes indicates a similar variable threshold. Meta has said this approach "helps ensure that notes reflect a range of perspectives and reduces the risk of bias." If published, the note appears as a banner at the bottom of the original post, which users can click to read the full note and supporting link.

In response to questions about how the algorithm assesses past disagreement between contributors, Meta clarified that, following X's open-source algorithm, the company "assesses differences in viewpoints entirely based on how contributors have rated notes in the past" and does not consider other engagement behaviors such as follows, likes or reposts. The algorithm "does not make assumptions about the nature of disagreement, such as political connotations of their agreements or disagreements (as it lacks data on political viewpoints or the subject matter of each note)." A technical expert who consulted with the Board suggested the following example to illustrate this point: it is possible that a country has a prominent liberal-conservative political axis, but that community notes contributors in that country rate notes according to whether or not the underlying content includes memes, supports a particular political party's agenda or supports a particular soccer star. In that case, the community notes algorithm will learn an axis of conflict that can best be described as capturing one of those divides, because that is the axis that best explains the data.

Users and community notes contributors can identify content potentially benefitting from a note, and contributors may write and propose notes on that content. Once a note is proposed, it appears in a feed to be voted on for helpfulness.

To facilitate rating of proposed notes, contributors have access to a feed of notes that still need ratings prior to being published, and which is accessible in the settings of the Facebook, Instagram and Threads apps. In response to a question about this feed, Meta clarified that notes are prioritized "based on recency and number of views on the post, in order to improve the timeliness and visibility of notes. The type of content is not considered when ranking notes (e.g., whether the content is civic or political). Notes with a higher helpfulness score may appear more prominently in this feed, whereas notes with a lower helpfulness score may appear less prominently." Moreover, Meta shared that it does not independently add posts to this feed nor "employ any levers to signal high-priority content in need of a note."

ii. Operational Aspects – Contributors and Policies

In its request, Meta describes how the community notes program works operationally. Unlike in Meta's fact-checking program, where the company partners with professional fact-checking organizations to review and label content, in community notes, Meta users apply to contribute to the program. Meta requires contributors to be over 18 years old, have an account that is more than six months old and that has not violated Meta's policies meant to "prevent the most severe harms – such as terrorism, child sexual exploitation, and fraud and scams," and have a verified phone number or have set up two-factor authentication. Should they meet these eligibility criteria, people are "gradually and randomly" admitted from the waitlist and may then write and rate notes. At present, contributors can compose and submit notes to "add more context" to public, organic content on Facebook, Instagram and Threads originating in the United States. Contributors must include a link supporting the context shared in the note. When rating notes, contributors have the option to state those written by other contributors are “helpful” or “not helpful” and explain their response by selecting a reason from a list of options, such as “whether the note is relevant to the post, easy to understand and uses neutral or unbiased language.”

The community notes product presently functions in six languages: English, Spanish, Chinese, Vietnamese, French and Portuguese. Meta has not yet decided “which languages will be supported in community notes for multilingual countries,” nor has it extended “the technical infrastructure of community notes to support writing, rating and publishing in languages beyond the six available at launch in the United States.” Meta shared that it expects to begin this work in early 2026. In response to a question about anticipated preparations for the international rollout of community notes, Meta said that it “intends to conduct testing on the community notes product prior to launch in a market. This testing will assess how the bridging algorithm is functioning in the local context and help identify any potential issues related to language representation or contributor participation.”

Meta said that notes are subject to the same Community Standards as organic content, with some significant enforcement differences. The company uses a “combination of automated systems, specialized review teams and expert manual review to assess whether notes violate our Community Standards.” Additionally, Meta reported that a machine learning classifier is also used to “reduce the distribution of posts people have requested notes on that our systems predict likely violate the Community Standards (but that have not been confirmed to violate) within the candidate posts displayed in the contributor interface while the content is awaiting human review.” Meta said this helps to reduce exposure to likely violating content.

According to Meta, contributor anonymity plays a role in community notes’ functioning. Contributor names are not displayed on notes they have authored or rated to ensure that ratings are based on “content and helpfulness, rather than the identity of who wrote them.” Meta said that this anonymity among contributors “encourages participation by mitigating peer pressure or fears of harassment.”

In its request, Meta said it is currently working on ways to keep contributors engaged, improve note quality and prevent coordinated abuse and manipulation of community notes. New contributors receive an in-product tutorial on how to write and rate notes. Active contributors can also join a community forum, hosted on Discord, to receive updates on new features and provide feedback to inform product development. Meta said that requiring contributors to possess a unique phone number and minimum account age will help to deter the creation of fake accounts. Gradually and randomly admitting users from the waitlist will “help protect against coordinated efforts to add notes to particular pieces of content or rate notes in a particular way." Because a note requires a link to information supporting the proposed context, Meta has instituted measures that block "high-severity violating URL links" to prevent abuse and encourage contributors to share higher-quality information.

IV. Research and Assessments of Community Notes-Style Moderation Systems

As part of its research and stakeholder engagement, the Board received a range of information assessing the opportunities and risks of crowdsourced, community notes-style moderation systems, as well as feedback, observations and empirical data on the performance of these systems on other social media platforms, but mainly X's community notes. In the absence of substantive data on the performance of Meta's system, the Board has extrapolated these findings from X's community notes to inform its recommendations in the latter portions of the opinion. This extrapolation may have shortcomings, as Meta's platforms differ from X's in several ways, including size, distribution of user base and virality dynamics. As more data about the performance of Meta's community notes becomes available, the Board's analysis based largely on X's community notes would need to be updated and tested further.

Community notes and the underlying bridging technology have shown promise in some respects. It has been argued that, in theory, it is a "promising approach to helping counter-speech combat misinformation" as it is "respectful of users' autonomy and encourages broad participation and healthy exchanges." Other feedback received by the Board cited studies that suggest community-sourced annotations on content can increase user confidence in moderation outcomes. Almost universally, stakeholders in the Board's consultations voiced support for approaches to deceptive information that prioritize freedom of expression, such as providing additional context through tools like community notes. They stressed that fact-checking and community notes should not be seen as mutually exclusive tools. Many highlighted how they could complement each other.

Based on some empirical research, community notes, if implemented with sufficient scale, speed, transparency and safeguards against manipulation, can function as an important tool for improving online discourse. Experimental and observational studies, as well as stakeholder contributions, argue that community notes can outperform generic warning labels in perceived legitimacy and user comprehension, and that they reintroduce a form of structured citizen participation into platform governance, grounded not in majority rule, but in cross-perspective, explanation-based collaboration.

There is also some optimism about the potential scalability and efficacy of crowdsourced notes, particularly when affixed to potentially misleading information. Some research has shown that, in aggregate, the judgements of laypeople can approximate professional fact-checkers' accuracy levels. This research argues that leveraging aggregate layperson evaluations through programs like community notes can "allow for scalable action against misinformation." Some studies have found that community notes can reduce engagement and reach of content labeled false or misleading, sometimes substantially, once a note is attached and seen by users. However, there are some divergences on these findings. Other research has analyzed temporal dimensions of engagement with misleading content and found that X's community notes "might be too slow to effectively reduce engagement with misinformation in the early (and most viral) stage of diffusion." Another research study found that the perceived social influence of the author of a source tweet affected levels of user consensus, with influential authors "yield[ing] a lower level of consensus among users" and fact-checking notes on their posts "more likely to be seen as being incorrect." Such disagreements likely stem in part from community notes' varied performance under different implementation conditions.

The Board also considered the views of stakeholders who argued that the performance of community notes programs in real-world conditions can be unpredictable. Systems that rely on algorithm-enabled contributor voting to append notes to content face challenges around coverage, participation, timeliness and reliability, all of which take on heightened importance in high-stakes situations like conflict and crises. At the same time, research on participatory and human moderation systems suggests that a degree of variability is inherent to such approaches, including expert fact-checking, and that the relative transparency of community notes may make this variability more observable and therefore more amenable to evaluation and iterative improvement.

Some research has found that, because community notes relies on contributors choosing which posts to annotate, its outputs reflect patterns of self-selection. Studies have shown that X contributors in the United States tend to write notes on counter-partisan posts and rate notes by co-partisans as helpful, potentially reinforcing ideological divides. These findings suggest that requiring consensus among those who usually disagree may be an incomplete solution to the problem of ideologically motivated notes, particularly in polarized online environments. However, other studies indicate that this cross-perspective consensus requirement also functions as a quality threshold that limits the visibility of ideologically motivated or weakly sourced notes, even if this design choice constrains overall coverage. Other reports and research have documented that, among published notes on X's community notes, professional fact-checkers are among the most-cited sources for supporting context, indicating the dependence of community notes on third-party fact-checking despite their distinct functions and aims. Therefore, reducing support for fact-checkers will likely impact community notes, as contributors may have fewer reliable sources to cite. This impact may be greater where there are limited reliable media sources. This interdependence underscores the importance of layered governance approaches, rather than reliance on any single mechanism, particularly in high-risk or time-sensitive contexts.

The relatively small number of notes that reach consensus and are published was highlighted as a central challenge by multiple stakeholders. According to one study, for 74% of misleading posts on X that had accurate notes proposed for them, those notes were never shown to users. Media investigations have found that most users never encounter published notes. When notes are published, they tend to receive lower exposure than the original post received before the note was affixed. This suggests users who see content that eventually receives a note often do not see the note when it is published. However, low publication and visibility rates may also reflect the system's intentionally high consensus and sourcing thresholds, which are designed to prioritize note quality and perceived legitimacy over volume. Disparities in the number of notes proposed and published across different language groups have also raised concerns about adequate representation of different linguistic groups in the community notes program. Such disparities may be influenced by contributor density, language-specific sourcing and platform investment in localization, rather than by the underlying viability of the model itself.

Some stakeholders also raised concerns that backlogs of unpublished notes on X (that is, notes that have been proposed by contributors but that have not yet reached the threshold of consensus necessary for publication) are negatively impacting the timeliness of community notes. One recent temporal analysis of publication patterns on X found that "notes, on average, are published 65.7 hours after the original post, with longer delays significantly reducing the likelihood of consensus." Another study found that, between January 2021 and January 2025, 87.7% of proposed notes on X remained in the "Needs More Ratings" category. Of all notes proposed for contributor rating, only 8.3% ultimately achieved "helpful" status and were published, with an average delay of 26 hours – "well past the point of peak visibility for most misleading posts." In a September 2025 update, Meta reported similar figures, with only about 6% of notes ever being published.

Analysis of trends on X has shown how low publication rates can have impacts on contributor engagement and retention. A recent study found that "most notes are produced by a minority of contributors" and "the fraction of authors who remain in the program year-on-year has slowed down since 2023." Moreover, the "fraction of authors who remain active is decreasing, suggesting a reduced capacity" to hold on to contributors. According to these authors, the repeated experience of not having a note published disincentivizes contributors from continuing to participate. The trend "speaks to the risks to the sustainability of the community notes system." If, as its public statements and policies suggest, Meta envisions community notes as a part of its "approach to misinformation," then these trends suggest community notes could result in too few good notes for the system to be a meaningful check on deceptive information.

The implications of these research findings in crisis and conflict situations were flagged as an area of concern. Multiple stakeholders shared research regarding X's community notes that challenges its ability to effectively address misleading information, particularly in moments where information integrity is of heightened importance, such as elections, crises and conflicts. For example, research into posts on X from five high-profile accounts that pushed false information during the 2024 Southport riots in the UK found that these accounts amassed over 430 million views. According to this analysis, of the 1,060 posts shared by these accounts during the height of the riots, only one received a community note. This data raises serious concerns about the ability of community notes to respond quickly to misinformation in crisis situations, particularly in the early, viral stages of its diffusion.

Some stakeholders raised concerns about X's recent decision to allow AI-powered contributors to "increase the number of notes" (see public comments, e.g., PC-31568, Full Fact). The Board sees risks in deploying AI-powered contributors to generate notes. The proliferation of AI-tools introduces new vectors for gaming and manipulation. Malicious actors could fine-tune models to subtly favor narratives, selectively frame evidence or exploit the rating mechanism – all while appearing neutral. Because most AI systems are trained on overlapping data and optimized for similar objectives, their participation could homogenize reasoning styles and interpretations in notes, crowding out minority and non-centrist perspectives. Meta told the Board that it "does not plan to allow AI note writers (i.e., AI-powered chatbots or agents) to submit community notes on Meta's platforms. Contributors may use AI to help them write notes; however, a human must submit the note under their name, and it is the human being who is considered the note's author."

Several experts consulted by the Board, as well as external stakeholders, raised concerns about the global applicability of X's community notes algorithm, which was designed to model and "bridge" political disagreement primarily along on a single axis, to societies where multiple axes of meaningful division exist. Deploying the algorithm in these settings potentially reinforces a simplistic understanding of complex political conflicts and social divides. However, some experts noted that the algorithm was not designed to capture all dimensions of political disagreement, but rather to identify limited areas of cross-perspective agreement, which may still provide meaningful contextual information even in societies characterized by multiple and overlapping axes of division. Practically, the algorithm not capturing social divisions could result in the system generating notes that tend to address a subset of expression, which may not correspond to the most salient, polarizing topics in a country. For example, some research shared by stakeholders found that "roughly half" of notes were applied to what they called "soft news" (stories covering culture, lifestyle or lighter topics rather than politics or current events). By comparison, "well-sourced notes surrounding politically charged events frequently went unpublished" (PC-31597, Institute for Strategic Dialogue), presumably because they had not reached the necessary threshold of consensus. Alternatively, this pattern could reflect publication thresholds intended to prioritize legitimacy and prevent capture, rather than the absence of accurate or well-sourced contributions, and that such selectivity is a common feature of participatory governance systems. Finally, the system could fail to surface minority or non-centrist perspectives, no matter their accuracy, particularly if the algorithm fails to capture socially and politically significant divisions beyond a presumed primary axis of polarization in a country. The human rights implications of these dynamics are considered below.

Taken together, the evidence and stakeholder input reviewed in this section indicate that community notes presents both significant limitations and potential strengths as a mechanism for addressing potentially misleading content. The challenges identified, relating to scale, timeliness, consensus, thresholds, contributor incentives, language disparities and the modeling of disagreement, underscore the importance of careful design and implementation choices. At the same time, research and consultations suggest that these constraints reflect trade-offs inherent to participatory, explanatory moderation systems. As such, their role should be assessed in light of their capacity to function as a line of intervention that prioritizes transparency, user participation and freedom of expression.

V. Human Rights Analysis

In 2021, Meta announced a Corporate Human Rights Policy, in which it outlined its commitment to respect rights in accordance with the United Nations Guiding Principles on Business and Human Rights (UNGPs). The UNGPs, which were endorsed by the United Nations (UN) Human Rights Council in 2011, establish a voluntary framework for the human rights responsibilities of businesses. These rights include, “at minimum … those expressed in the International Bill of Human Rights” (Principle 12).

As a global corporation committed to the UNGPs, Meta should respect international human rights standards wherever it operates (Principle 11), and “seek to prevent or mitigate adverse human rights impacts that are directly linked to their operations, products or services by their business principles, even if they have not contributed to those impacts” (Principle 13). The UNGPs also establish that businesses should carry out human rights due diligence to assess actual and potential impacts and act upon their findings (Principle 17). To do that effectively, businesses should monitor qualitative and quantitative indicators and incorporate input from impacted stakeholders (Principle 20).

Through its cases and advisory opinions, the Board assesses the human rights impacts of policy and enforcement decisions. When these cases reveal that Meta is causing a negative impact, or may not be taking steps to identify, monitor and limit negative outcomes more broadly, the Board makes appropriate recommendations. In a policy advisory opinion, the Board focuses directly on Meta’s policy and enforcement choices to assess whether the company is upholding its commitment to respect rights under the UNGPs.

Applied to community notes, the Board explored whether the program in practice serves to address and mitigate potential adverse human rights impacts according to Meta’s responsibilities, especially in markets outside the United States where Meta may roll out community notes. In the sections that follow, the Board first focuses on contexts where community notes could directly create or heighten the risk of harm to users. These affirmative harms are distinguished from the potential inadequacy of community notes as a harm mitigation measure (particularly in relation to deceptive information that falls short of Meta’s threshold for removal), which is addressed in the subsequent section.

Direct Impacts to Human Rights from Deploying Community Notes

Principle 13 of the UNGPs requires Meta to “avoid causing or contributing to adverse human rights impacts” and to “address such impacts when they occur.” A wide array of rights may be impacted by the community notes program. For example, freedom of expression, which includes the right to seek and receive information (Article 19, International Covenant of Civil and Political Rights (ICCPR); General Comment 34, 2011, para. 11), may be enhanced to the extent that community notes serves as a vehicle for contributors to have additional tools for their expression, foster counter-speech or generate and make accessible contextual information that might otherwise be unavailable to users. Debate, exchange of ideas and collective decision-making are central features of Meta's vision for the system's functioning. Content receiving a community note is not demoted, as can be the case with content that receives a fact-checking label. In this regard, community notes could be viewed as further contributing to the expressive rights of some platform users.

On the other hand, the Board finds that introducing community notes in countries outside the United States could, in certain circumstances, pose significant human rights risks and may generate or contribute to tangible harms that Meta has assumed a responsibility to avoid or remedy. While the research and stakeholder observations on X's community notes shared above give some sense of these harms, their likelihood and severity are largely dependent on country-level context, as well as Meta's plans for expansion. Meta shared information on this topic in its request and answers to the Board's questions, including its plans to test the community notes product prior to launch in a market. This testing will aim to "assess how the bridging algorithm is functioning in the local context and help identify any potential issues related to language representation or contributor participation." In light of UN Office of the High Commissioner for Human Rights resources for implementing the UNGPs in the technology space, Meta should focus this testing on "understanding how the local socioeconomic, political and human rights realities either exacerbate or protect against human rights harms" as the global rollout of community notes approaches. The Board urges Meta to publicly report on the results of that testing prior to rollout.

Although Meta frames community notes as additional "context," it could create chilling effects on freedom of expression, particularly where users fear that increased scrutiny than can come with public labeling of their speech could lead to retaliation or potential harassment, or fear their notes are traceable. This may lead users to avoid posting on controversial or sensitive topics altogether. The Board notes, however, that this risk is not unique to community notes, and is likely to be present in any system designed to counter misinformation. In political contexts where the civic space is constrained and human rights are at risk, community notes may discourage dissent, citizen journalism and user-led fact-checking, or minority viewpoints, especially in the absence of transparency about how the system works or meaningful avenues for appeal. Meta shared that it is exploring adding the ability to request additional ratings from contributors as a potential remediation path. Currently, users do not have the ability to appeal a community note, as is available for content that receives a fact-check label, if, for example, they believe their intended meaning has been changed or distorted by a note or if a note itself is inaccurate. Meta's development of this feature could serve to mitigate the risk of a chilling effect.

In contexts where governments suppress dissent, including through attacks on freedom of expression, the system may put individuals at risk of retaliation if anonymity is compromised. This creates risks to contributors' rights to privacy (Article 17, ICCPR), to security of the person (Article 9, ICCPR) and even the right to life (Article 6, ICCPR). Even when formal anonymity is provided, participation in community notes could expose contributors to surveillance or targeting if identities or behavior patterns are inferred. Data about who contributes to or interacts with notes may reveal political opinions or affiliations, placing contributors at risk in countries with histories of using such information for repression. Meta said that its primary mitigation around disclosing contributor identities in community notes is the design of the data model itself. Contributor entities, which are used to create and rate notes, are separate from Meta user profiles. However, in the absence of a more detailed account of the technical architecture of how Meta has implemented contributor accounts for community notes, the Board cannot make firm conclusions about whether this risk will be adequately mitigated by Meta's current plans. In advance of rollout to contexts where governments suppress dissent, Meta should ensure its protections are robust through red-teaming of community notes' technical architecture under adversarial conditions, such as threat modeling for state actors.

The Board is also concerned about the potential use of community notes by coordinated disinformation networks, especially in environments with past or present evidence of intentional, large-scale manipulation of or attacks against information ecosystems. Social media platforms, including Meta, are aware of this risk and have disclosed several instances when they have removed accounts for such behavior. In these situations, malicious actors could coordinate to "game" or manipulate participation and rating mechanisms, allowing them to, for example, systematically discredit opposing views or publish notes that do not reflect actual consensus among contributors. Such abuses of community notes, if successful, could manipulate public discourse in ways that skew information environments and impact people's right to participation in the conduct of public affairs (Article 25, ICCPR). Meta detailed several safeguards it has in place to prevent gaming of community notes. It said that it "monitors for potential gaming of the community notes program by analyzing patterns of contributor behavior and interactions." The company said that, to date, it has not detected any coordinated inauthentic behavior or gaming of the program.

However, recent research on X's community notes found that "a small minority (5-20%) of bad raters can strategically suppress targeted helpful notes," thus raising concerns about the adequacy of safeguards and, in turn, the systems' vulnerability to manipulation. The published status of Meta's notes does not "lock" until two weeks after consensus is reached and the note appears. This could make community notes vulnerable to malicious actors who could coordinate to remove a note by continuing to rate it unfavorably in the two weeks between it being published and locking. Moreover, these vulnerabilities are likely to be heightened with the proliferation of AI tools, and particularly agentic AI, that allow for easier scaling of account creation and management. While Meta's safeguards are, in theory, meant to monitor for such activity, it is not clear from the information provided to the Board that these safeguards adequately address the potential scale of the threat. Further assessments of both the severity of this risk and the adequacy of Meta's safeguards will depend on empirical data examining whether AI tools used in community notes create or exacerbate vulnerabilities such as coordinated gaming or scaled information operations, which Meta will need to collect and evaluate.

Community notes can also risk privileging dominant political, ethnic or linguistic groups and potentially marginalizing disfavored minorities, particularly if multiple groups share prejudice towards a minority group. This problem is reinforced if such groups are not represented or are underrepresented as contributors, potentially resulting in the community notes algorithm modeling disagreements that are not illustrative of critical societal divisions. This raises concerns about potential impacts to the rights of members of such groups, for instance, the rights to equality and non-discrimination (Articles 2 and 26, ICCPR) and freedom of expression (Article 19, ICCPR), especially if objective circumstances make them less likely to participate in the creation and rating of community notes. There are several factors that could make this harm more likely or more severe. As discussed by experts and stakeholders, outlined above, one major challenge for X's community notes system has been applying an algorithm that models polarization and social division along a single axis to contexts where that is not the case. Stakeholders consulted by the Board noted that such a blunt approach could result in the system skewing information environments in favor of majoritarian perspectives and potentially publishing harmful notes that target mutually disfavored minorities. These dynamics could potentially lead to increased division.

In these cases, shared prejudice towards disadvantaged minority groups could serve as the "bridge" between other majority groups that otherwise disagree on other major axes of polarization. This type of harm can also become more likely if disfavored minorities have differential access to the internet and social media and are thus underrepresented as contributors in community notes. For community notes to enable broad access to plural and diverse sources of information as Meta envisions, robust participation across groups, but particularly relevant language groups and other minority groups (i.e., by religion, ethnicity, gender, age), is essential. For example, a consortium of South Asian non-governmental organizations (NOGs) presented some evidence of harms stemming from a majoritarian dynamic on X's community notes in Indian contexts, where political divisions reflect complex and overlapping affiliations spanning ethnicity, religion, language and caste. While it is difficult to know how generalizable the problem may be based on this limited example, the risk is sufficiently plausible to warrant Meta's careful assessment of social division and the position of disfavored minorities in markets where community notes may be introduced.

Experts in bridging algorithms consulted by the Board pointed out that there is no technical reason why the community notes algorithm could not be designed in a more sophisticated way to take multiple axes of disagreement into account at the same time. As noted above, in the absence of further design information, the Board is assessing the risks here on the basis of how the X algorithm, which Meta has taken as its point of departure, has been shown to operate. Insofar as Meta designs its version of community notes to function in a more multidimensional way, that could be a means to mitigate the risks identified here.

The severity and likelihood of the different harms referred to above can depend significantly on whether community notes is complemented by other measures to address deceptive information, such as third-party fact-checking, especially given the dependence of community notes on their work, as noted previously. Conversely, other external factors, such as robust media independence and widespread digital literacy, could reduce the likelihood and severity of these harms.

Indirect Impacts on Human Rights from Deploying Community Notes

The most directly applicable human rights responsibilities related to community notes that Meta has under the UNGPs are those explored above. These address the conditions under which community notes could, in their own right, cause or contribute to adverse human rights impacts under certain local circumstances. In addition to those direct responsibilities, Meta's global rollout and use of community notes also implicate broader human rights concerns that are more indirect in nature, but nonetheless significant.

Meta has said it will continue to enforce the misinformation and harm policy lines, prohibiting misinformation that is: i) likely to directly contribute to a risk of imminent violence; and, ii) likely to directly contribute to a risk of interference with people's ability to participate in elections. However, users also post false and misleading information that can, in a variety of circumstances, lead to conditions in which human rights might be at risk, but where the risk falls short of that threshold for removal. For example, the UN Special Rapporteur on freedom of expression has said misinformation "is politically polarizing, hinders people from meaningfully exercising their human rights and destroys their trust in government and institutions" (A/HRC/47/25, para. 2). It may contribute to a broader set of harms, "chilling free speech, reducing levels of trust in the public sphere as a space for democratic deliberation, amplifying anti-democratic narratives, driving polarization and promoting authoritarian and populist agendas" (A/HRC/47/25, para. 24). A degraded information environment in which such conditions exist can thus put people at risk by limiting their access to information and their right to participate in the political process.

While these harms may be serious, they typically fall below the threshold that would justify state restrictions on freedom of expression under Article 19(3) of the ICCPR. Indeed, the UN Special Rapporteur on freedom of expression has raised concerns (A/HRC/47/25, paras. 57-58) about the risks to freedom of expression posed by overly broad "fake news" legislation. Around the world, these bills have been used as the pretext for censorship of NGOs, journalists and political dissidents. In light of these concerns, lesser measures short of removal can be more proportionate to the legitimate human rights aims. Previously, the Board has emphasized the importance of addressing harms associated with deceptive information through measures less intrusive than removal, such as providing information to correct falsehoods and demotion of misleading content (see Posts Supporting UK Riots and Protest Footage Paired with Pro-Duterte Chants decisions). As the Board has noted repeatedly, these are different from state obligations in this area, and states pursuing similar restrictions would infringe on rights.

There is some ambiguity on Meta's part about whether the company intends community notes to be a mitigation measure for these types of harms. On one hand, in response to the Board's question about the relationship between community notes and third-party fact-checking, Meta indicated "fundamental differences" between the two, saying that community notes serves a broader objective: contributors "add context to any piece of content where they believe additional information may be helpful." On the other hand, Meta introduced community notes in the same statement that announced it was ending its third-party fact-checking program in the United States. The company describes community notes specifically as a feature that lets people "add more context to *posts that are potentially misleading or confusing*" (emphasis added). Meta also lists community notes as part of its approach to misinformation, with third-party fact-checking operating as the “inform” mechanism outside the United States, and community notes performing the same function inside the United States.

The Board finds that these latter statements, and Meta’s misinformation policy more broadly, shows that the company affirmatively seeks to address this type of deceptive information, and that it envisions community notes playing a role in fulfilling that responsibility.

To the extent that community notes is intended to function as a tool to address human rights concerns related to misinformation that does not meet the threshold for removal, the Board welcomes that community notes do not restrict speech in the way other measures, such as content take-downs and account strikes, do. Content receiving a note does not experience reduced visibility or reach, unlike posts receiving labels resulting from third-party fact-checking. The Board finds this an appropriate way to modulate the proportionality of Meta’s intervention to make it more consistent with the protection of freedom of expression. In some contexts, the program can also generate counter-speech that serves as an effective check on this type of misinformation and therefore help Meta meet its assumed responsibility in this area. However, should Meta envision community notes as its primary or only way to address all types of deceptive information not meeting the threshold for removal, design decisions inherent to its functioning may limit its ability to do that. The effects of these decisions, such as the algorithm’s imposition of a single axis of disagreement on complex societal divisions, delays in note publication, the relatively small number of notes published, and, above all, the system’s dependence on the depth and reliability of the broader information environment (including third-party fact-checking where available), raise serious concerns about the extent to which community notes is fit for the purpose of addressing deceptive information linked to harm.

Recently, the Special Rapporteur on freedom of expression expressed concerns about the human rights implications of some of these features, as well as the industry-wide shift to community-driven approaches to fact-checking, like community notes (A/80/341, paras. 67-72). While noting the potential benefits of community-driven moderation, including scalability, cost and its democratic nature, the Special Rapporteur also flagged potential weaknesses. These include the system’s “susceptibility to capture” through manipulation of ratings, “inconsistent application of standards,” and “lack of consistent expertise and requirements of community consensus” (A/80/341, paras. 69-70). The Special Rapporteur concludes that “community notes and fact-checking should not be seen as mutually exclusive tools” (A/80/341, para. 72). As discussed above, in deeply polarized contexts where social divisions do not map neatly onto the model of polarization assumed by the community notes algorithm, consensus ratings can be difficult to achieve. Practically, this technical limitation can result in too few notes being shown for the system to be a meaningful check on deceptive information. Moreover, Meta's practice of using its automated technology to reduce the distribution of borderline posts (content that its systems predict likely violates the community standards but has not been confirmed as such) on which people have requested notes in the contributor interface may cause those posts, which most need contextual interventions, to be less likely to receive notes. Beyond the functioning of the algorithm, the Board notes Meta's practice of reducing distribution in the contributor interface as an additional explanation for why community notes may show too few notes to be a meaningful check on deceptive information. The Board also recognizes, however, that the risks of capture, inconsistency and uneven expertise are not unique to community-driven systems, and may also arise in expert-led or automated moderation.

While some research has found that community notes are perceived as more trustworthy than misinformation flags from fact-checkers, other research on the accuracy of community notes has found that crowd-driven evaluations are not as accurate relative to professional fact-checkers. Moreover, it is important to note the role of sources provided by professional fact-checkers in community notes that reach helpful status. One recent research study found that "community notes cite fact-checking sources up to five times more than previously reported," while another study found that "fact-checking organizations are the third most used reference globally." These comparisons do not imply that Meta's fact-checking program is without significant limitations and weaknesses as well. For example, the Board has previously found that the overwhelming majority of content in the queues for fact-checking is never reviewed by fact-checkers, in part due to operational and resourcing constraints (see Removal of COVID-19 Misinformation policy advisory opinion). Thus, even though the limitations of community notes do raise legitimate concerns, the Board observes that measures based on contextual counter-speech may, in certain circumstances, better satisfy proportionality and freedom of expression considerations than more intrusive interventions, particularly where misinformation does not meet the threshold for removal.

More importantly, insofar as the community notes system is intended to serve as part of Meta's comprehensive response to human rights concerns arising from false and misleading information that falls short of likely and imminent harm, its deployment cannot be assessed in isolation. In some markets, languages and issue areas, professional fact-checking capacity is limited or absent, and community notes could be a means by which users receive contextual information about potentially misleading content. Elsewhere, more robust third-party fact-checking operations could be supplemented by community notes (see PC-31587, European Fact-Checking Standards Network). Meta should consider the overall condition of an information ecosystem – including the accessibility and reliability of information – when determining where and how to introduce community notes. Given the role that fact-checking agencies often play in shaping broader information environments through research, sourcing and reporting, their presence (or absence) should be one factor that is considered by Meta when making these determinations. This consideration is particularly important in conflict-affected settings, during elections and in environments where historically marginalized groups and disfavored minorities lack adequate access to participation in public discourse.

Human Rights Due Diligence

Meta described its due diligence in advance of the U.S. launch of community notes as consisting of "review of academic literature and studies" focused on X's version of community notes, as well as bridging algorithms more broadly. Meta said that it conducted this review, which focused solely on the U.S., to assess various factors that could be considered indicators of the program's effectiveness. The review included examining research on whether the U.S. public perceives community notes as more legitimate than traditional fact-checking, which could suggest a "greater likelihood of people accepting information provided by a community note over that from a third-party fact-checker." Additionally, in response to questions from the Board, Meta said it "evaluated studies addressing the accuracy of community notes, exploring whether they can match or even surpass the accuracy of third-party fact-checking." Meta did not indicate that any of this research focused on human rights harms related to community notes. Meta said that its human rights policy team is now advising relevant cross-functional teams to identify potential human rights risks and benefits of community notes. This advisory opinion should also be considered part of the human rights due diligence around Meta's global launch of community notes.

The Board's recommendations and associated analysis of the factors Meta should consider when omitting countries from the community notes rollout have been informed by a broad consideration of the potential human rights impacts of an international rollout presented above. However, the potential human rights risk and appropriate mitigation strategies depend critically on the design and functionality of the community notes product in each context. Meta has claimed sufficient mitigations of human rights risks (on, for example, contributor anonymity and protections against gaming) in materials provided to the Board, but the effectiveness of these cannot be adequately verified unless and until there is data and reporting on how they function in practice.

The Board's recommendations are therefore necessarily conditional, as the questions posed by Meta cannot be answered with a high degree of confidence in the absence of detailed data about how Meta's community notes algorithm functions in real-world situations, as well as its relationship to other misinformation tools. The company will also need to carefully monitor and measure the actual impacts as the product becomes operational in new and different contexts. Meta has committed to testing the community notes product prior to launch in a market. The Board notes that this testing should focus on surfacing and mitigating risks related to the functioning of the algorithm, language representation and contributor participation. The Board underscores its repeated call for substantial transparency, reporting and researcher access to data on community notes performance. This should include equivalent data to what X currently shares about its community notes system, including, at a minimum, the full text of all notes and their associated geographies, ratings for each note, status history, user enrolment and the requests for notes on content.

VI. Recommended Country-Level Factors

The Board recommends the following factors to guide Meta's rollout of community notes. To assess these factors and weigh them against each other in a scalable manner, the Board also suggests relevant rankings and assessments that Meta should review as part of its preparations and eventual monitoring of community notes' expansion. These rankings and assessments are not comprehensive sources of truth about country-level conditions but give general measurements of some of the factors below. Meta should also review its own data about community notes' performance and consult with experts and impacted stakeholders from the countries that may receive the program.

As discussed above, it is possible for Meta to do a human rights compliant rollout of community notes. Several factors may impact Meta's order of implementation, and some factors require additional due diligence. There are also other factors that the Board views as threshold considerations where Meta should ensure particular mitigations are robust before proceeding.

Repressive Human Rights Contexts

As discussed throughout, the community notes program depends on an active and engaged contributor base. Community notes-style systems, and bridging technologies more broadly, have experienced their greatest successes in contexts characterized by a robust civil society. Ideally, contributors have access to independent media to support proposed community notes and participate without fear of harassment or retribution for writing and rating notes. These conditions may be more difficult to achieve in countries with repressive human rights records. In such environments, state influence, coordinated online actors and fear of retaliation can skew community participation, leading to the over-representation of majoritarian or state-aligned narratives in the contributor pool. This can produce chilling effects that pre-emptively silence opposition voices within the system.

On the other hand, the Board also considered how environments without indicators of robust civil society, such as low levels of freedom of expression and historical and current evidence of systemic human rights abuses, might benefit from the introduction of community notes (see PC-31548, Cubalex). Here, if robust contributor protections are ensured, community notes could empower contributors to challenge repression by using the system to surface alternative perspectives and publish information that might otherwise be inaccessible. If anonymity is preserved, the system could also be a venue for challenging official state narratives.

Because of potential risks to the safety and security of contributors, the Board recommends that, until Meta can demonstrate that contributor privacy protections are robust and effective with evidence of red-teaming under adversarial conditions, a clear policy on handling requests for community notes data from law enforcement agencies and the presence of risk mitigation measures, countries with repressive human rights records and weak civil societies be omitted from the initial international rollout of community notes. This is a necessary condition prior to expansion for the Board. Rankings like Freedom House's Freedom in the World and Freedom on the Net, as well as the Varieties of Democracy (V-Dem) project, Article 19's Global Expression Report and Reporters Without Borders' Press Freedom Index, employ several indicators (beyond overall country scores) that offer holistic measurements of levels of freedom of expression, press freedom and civil society engagement.

High-Risk Elections

The Board recognizes that some electoral contexts may be appropriate for community notes. Where information environments are robust, media freedom allows for journalists to report multiple perspectives on relevant issues and civil society actors operate without the fear of retaliation, community notes can support access to information and freedom of expression in electoral contexts. However, where those conditions are not present and the system risks publishing notes that mislead people about facts essential to their exercise of the right to take part in the conduct of public affairs at critical moments in electoral processes, Meta should proceed with caution. Where these human rights risks are present, and Meta determines through product testing, risk assessment and human rights due diligence that its safeguards are insufficient to mitigate the risks, community notes should not be introduced in advance of or during major elections. For the Board, this is a threshold that Meta should use to determine if, at any given moment, community notes is ready to be introduced to a country.

As discussed above, some research suggests that the speed of misinformation can far outpace the capacity of community-based moderation systems. For elections that are time-bound events characterized by intense political activity, delayed intervention can have immediate impacts on rights to political participation. Initial exposure to deceptive content may shape responses or undermine trust in institutions and political processes before notes offering alternative interpretations of content or corrections are surfaced. Elections that spill over into post-election periods because of contested results, for example, can be impacted by the persistence of deceptive narratives. In both these cases, community notes does not create affirmative harms (as it would in the case of surfacing hate speech, for example, or exposing contributors to retaliation from government authorities), but, when deployed by itself, is potentially an inadequate mitigation measure for misleading information that falls short of likely and imminent harm. For these reasons, the Board recommends that any introduction of community notes in advance of an election address how community notes is integrated with other election-related measures, such as Integrity Product Operations Centers and Election Operations Centers.

In developing possible expansion timelines, Meta should consult with relevant rankings and data pertaining to election integrity, as well as calendars that map upcoming global elections, such as the International Foundation for Electoral Systems' (IFES) Election Guide.

Crisis and Conflict Situations

In crisis and conflict settings, all of the risks discussed throughout may be present. Community notes may be vulnerable to coordinated manipulation by armed groups, state actors or their supporters seeking to legitimize propaganda through gaming of the note rating system. Notes that are published but reflect inauthentically manufactured consensus risk contributing to skewed information environments that put users at risk. Information asymmetries on community notes can be exacerbated when particular groups are unable to participate as contributors on account of factors that predominate in conflict settings, such as a lack of stable internet access and the safety needed to participate. Beyond intentional manipulation of information, false or misleading content related to violence and public safety that may not qualify under Meta's misinformation and harm policy lines can spread rapidly and have offline consequences before community notes are proposed, rated and published.

The system also risks publishing violent content if the community notes algorithm identifies a division that does not correlate with the drivers of violence in a particular conflict. In this scenario, the system could potentially "bridge" division between dominant groups that share hostility toward another group involved in the conflict, resulting in notes that target and potentially incite violence against this mutually disfavored group.

The consequences and human rights impacts of these potential issues are heightened in crisis and conflict scenarios because of the speed of critical developments. This makes the question of timeliness critical. The temporal lag in publishing notes, which was flagged by multiple stakeholders as a top concern, suggests that community notes may not be an adequate primary safeguard in crisis and conflict scenarios. Moreover, in conflict and some crisis situations, there is an existing propensity for violence, which means thresholds for incitement might be lower. As a result, notes that target particular groups can more easily result in offline harm. Here, the Board reiterates its recommendation in the Posts Supporting UK Riots decision for Meta to "undertake continuous assessments of the effectiveness of community notes as compared to third-party fact-checking," focusing on the "speed, accuracy and volume of notes or labels being affixed in situations where the rapid dissemination of false information creates risks to public safety."

As with elections, the Board presumes that other crisis- and conflict-related integrity measures, such as the Crisis Policy Protocol, would remain in place. Meta should ensure that any rollout of community notes that implicates such situations addresses how community notes is integrated with these crisis-related protocols and tools. The Board is concerned that, as Meta explained in its response to a question from the Board, the company "has not developed provisions regarding the use of the product in crisis situations, including adapting, modifying, or suspending the feature." Given concerns about the uneven performance of the community notes and the absence of internally verified pilot data about the system's performance during crises and conflicts, Meta should consult relevant rankings and data, such as the Armed Conflict Location and Event Data Project (ACLED), to avoid relying upon community notes during such events.

Because of this uncertainty about how community notes would perform in response to complex conflict dynamics, as well as the system potentially causing or heightening the risk of harm, the Board does not believe that community notes should be introduced in countries experiencing crises or protracted conflict. For the Board, this is another threshold condition that Meta should use to guide the introduction of community notes.

Language Complexity That Meta Cannot Technically and Operationally Accommodate

For community notes to function as Meta intends, the contributor base should reflect the different languages used in a given context. In countries where the program does not function in those languages, or Meta is unable to achieve the levels of linguistic representation and participation required for the system to function properly, Meta should delay introducing community notes. Not doing so could create or exacerbate language disparities in the notes that are proposed and published on community notes, thus undermining community notes as a source of plural and diverse information.

Moreover, the Board has repeatedly called on Meta to ensure linguistic parity, both technically and operationally, across its content moderation systems. These calls have focused on functions like developing automated enforcement tools that are attuned to context across languages, ensuring that classifiers are accurate across languages, working with Trusted Partners in different languages, and maintaining relationships with third-party fact-checkers who can evaluate claims in different languages within a given market. Community notes alone cannot fulfill all functions of Meta's content moderation.

Finally, the Board notes that there are potentially relevant linguistic and cultural variations in how different groups use community notes. For example, what it means to click the button that rates a note as "helpful" in different countries might differ in unexpected ways based on norms about how that button should be used, or the exact connotation with which the word "helpful" is understood in the community notes user interface when translated into different languages. These differences will change how the dataset that the community notes system accumulates is structured and the meaning that it should be attached to it – and ultimately the types of notes that it publishes. In countries where Meta anticipates these features of community notes cannot accommodate linguistic complexity, it should delay the introduction of the program until research and testing demonstrates that users across cultural and linguistic contexts are able to understand and engage with the system.

Where Social Division and Disagreement That Drives Violence Does Not Map Onto a Single Axis of Polarization

In general, the design of X's community notes algorithm models disagreement and division in a particular context along a single axis of polarization. It will attempt to "bridge" groups that exist on either end of that spectrum. Meta has not provided any information that suggests its program will be substantively different. However, in contexts where division and disagreement cannot be easily modeled along a single axis, this may not be an appropriate assumption. In these contexts, community notes will likely not address how multiple factors intersect (across politics, ethnicity, religion, language and caste, for example) to fuel disagreement.

In practice, this means that in some contexts, community notes may find consensus along an axis with limited relevance to the most critical societal divisions. In some scenarios, harmful notes that reach consensus may be published. This can result, for example, when dominant groups share a mutual prejudice against a minority group, and that prejudice serves as the "bridge" between those dominant groups. In contexts where the algorithm does not identify and attempt to "bridge" a division that drives conflict and violence, this risk of harm is especially acute. For these reasons, the Board recommends that Meta exercise extreme caution when considering countries characterized by these dynamics.

The Board recognizes that all sociopolitical contexts are characterized by multiple axes of division and disagreement to some extent. However, there are contexts where the risks of misalignment between how community notes models disagreement in a society and what drives social and political division in that society are heightened. This includes countries where linguistic, ethnic and religious fractionalization characterize political culture. Such countries need not be categorically omitted from Meta's plans to expand community notes. Rather, the global introduction of community notes should be sequenced in a way that allows for testing of the system's performance in different contexts and should begin cautiously in environments more similar to places where Meta already possesses data, testing and risk mitigation measures. The community notes algorithm could also be designed to model and integrate multiple axes of division at the same time. Here, Meta's rollout (and the temporary omission of some countries that do not meet the threshold of early introduction) would depend on design choices that might distinguish Meta's community notes from X's version and improve upon it.

The Varieties of Democracy (V-Dem) project produces annual rankings of various indicators of democracy. These include data on polarization levels in different countries, though they offer less insight into the nature of polarization in different countries. As part of its human rights due diligence related to the community notes global rollout, Meta should consult with experts on countries characterized by high polarization to assess whether its algorithm is capable of capturing salient social and political divisions.

History of Coordinated Disinformation Networks

Community notes operates under the assumption that a sufficiently diverse and independent set of contributors will evaluate content in good faith and that consensus signals can reliably approximate accuracy. In environments where state-aligned actors, commercial influence operations and other malicious actors have repeatedly demonstrated the ability to coordinate large numbers of accounts to promote deceptive information, this assumption may not hold.

Meta has told the Board that it believes its safeguards against the manipulation and gaming of community notes to be sufficient. Continued real-world testing of community notes' vulnerabilities will produce data necessary to verify whether those safeguards are adequate. If not, community notes risks becoming a vector for manipulation rather than a safeguard against it. This risk will only become more acute as artificial intelligence facilitates the scaled creation and operation of accounts and networks.

The Board recommends that Meta initially omit countries where intentional, large-scale disinformation networks have historically been based. In doing so, Meta should focus on markets with demonstrated high prevalence of coordinated inauthentic behavior and influence operations targeting domestic audiences, as evinced in Meta's own reports on these topics. Eventual inclusion in the community notes program should be conditional on the results of testing, including red-teaming of program vulnerabilities, showing that safeguards are adequate. Meta should consider not only the presence of malicious actors, including those linked to authoritarian regimes, but also whether such actors have demonstrated the intent to manipulate information ecosystems and possess the technical sophistication to do so on a large scale. Country-level reports in Freedom on the Net and the Press Freedom Index document instances of manipulation campaigns and state-sponsored disinformation, as do company disclosures on the topic.

Obstacles to Internet Access

The Board recommends omitting countries that face persistent or systemic obstacles to internet access, as community notes relies on broad, consistent and equitable contributor participation to function as intended. Where access is limited by infrastructure gaps, high costs, intermittent connectivity, regional disparities and, especially, government-imposed shutdowns, the pool of contributors able to participate is necessarily narrow. This undermines the core premise that the community notes system reflects a diverse and representative set of perspectives capable of collectively deciding what context is important and if it is relevant to potentially misleading claims. Moreover, if particular groups are excluded because of access issues or other groups are overrepresented, the system could risk bias amplification.

As Meta notes in its request, government restrictions on the internet are a particularly relevant sub-factor because "limited internet freedom (due to regulation or censorship) could limit people's access to timely and accurate information for notes and make it harder to identify reliable sources to link within their notes (e.g., restrictions on using particular search engines or accessing certain media websites)." Freedom on the Net offers a holistic measure of obstacles to internet access, more generally. This indicator encompasses infrastructural, economic, governmental, regulatory and legal obstacles to internet access.

Weighting

Throughout its recommendation of factors, the Board has noted where community notes could create or heighten the risk of harm to users. These affirmative harms have been distinguished from scenarios where the possibility of harm comes from the inadequacy of community notes as a harm mitigation measure, particularly in relation to deceptive information that falls short of Meta's threshold for removal. The Board recommends that the factors where affirmative harms are more likely should be weighted more heavily, and it has indicated that these involve thresholds where Meta should not introduce community notes until it demonstrates it can do so in a way that ensures the risks of these harms have been mitigated. Elsewhere, the Board has recommended that Meta proceed cautiously and stagger or delay rollouts until testing demonstrates that its safeguards are adequate. This does not imply that these factors are less important. The Board notes that this guidance cannot be absolute. In its deliberations, the Board considered scenarios where the presence of one factor might predominate such that omitting a country is appropriate, as well as scenarios where no one factor is particularly severe, but several factors compound in a way that makes omission appropriate.

Ultimately, Meta's human rights diligence and risk mitigation efforts should examine the factors recommended by the Board alongside internal platform data and external feedback on community notes performance. The Board notes that Meta already tracks some of the factors discussed above, as well as the rankings recommended as proxies for the factors, as part of its Integrity Country Prioritization Index. This is a biannual process that Meta uses to review and prioritize countries that have the highest risk of offline harm and violence. The Board expects that Meta will exercise the same degree of rigor and judgment in determining how the introduction of community notes will proceed in different countries.

In the analysis above, the Board has provided Meta with a series of factors it should consider in determining whether to omit a country from any future expansion of community notes beyond the United States, as well as guidance as to where the gathering and analysis of additional data on the program function will be critical to assessing the future relevance and weight of those factors. To demonstrate its consideration of the Board's recommended factors, Meta should provide the Board with the criteria or risk matrix it develops to guide expansion every six months during its period of initial expansion, along with evidence of how these are applied in country-level decisions about community notes' expansion.

The Board will consider this recommendation implemented when Meta submits its first report detailing how the outlined factors have been considered, balanced and weighted for each relevant market, and how its gathering and assessment of further data on program function will be integrated into the periodic reassessment of the appropriate weighting of the outlined factors.

***Procedural Note**

The Oversight Board's policy advisory opinions are prepared by panels of five Members and approved by a majority of the Board. Board decisions do not necessarily represent the personal views of all Members.

For this policy advisory opinion, independent research was commissioned on behalf of the Board. The Board was assisted by Duco Advisors, an advisory firm focusing on the intersection of geopolitics, trust and safety, and technology.