Meta’s Policy Advisory Opinion Request: Factors for Omitting Countries from Community Notes

Table of Contents:

I. Introduction

On January 7, 2025, Meta announced our decision to end the third party fact-checking program in the United States and transition to Community Notes. Community Notes is a new product that allows users to add more context to public, organic content on Facebook, Instagram, and Threads originating in the United States. Eligible users can sign up to become Community Notes contributors and, once admitted off the waitlist, can submit proposed Notes. A Note will be displayed on content only if a sufficient number of contributors, who usually disagree with each other based on past ratings, deem the Note "helpful".

When we launched Community Notes, we stated our intention to refine the product in the United States before expanding globally. As we evaluate our approach to international expansion, we are requesting a Policy Advisory Opinion ("PAO") on the factors we should consider when deciding which countries, if any, to omit from the international roll out. We are also requesting guidance on how to weigh those factors against one another in a scalable manner.

Since the product is still undergoing testing and refinement, its form may evolve. As such, we respectfully ask the Board to focus its examination on the country-level factors relevant to omitting countries from the international roll-out, and not on topics such as general product design or the operation of the Community Notes algorithm.

We appreciate the Board's independent expertise and thoughtful consideration of the issues presented in this PAO request, which will be helpful to our program's international expansion.

II. Overview of Community Notes

A. Eligibility and Baseline Criteria

Community Notes are written and rated by Community Notes contributors and not by Meta or a small group of fact-checkers. Any user can apply to become a contributor as long as they meet our eligibility criteria: (1) be over 18 years old; (2) possess an account in good standing[1] that is at least six months old; and (3) have a verified phone number or be enrolled in two-factor authentication. Contributors meeting those criteria are gradually and randomly admitted from the waitlist. Once

[1] An account is in good standing if it has no violations of content policies in areas such as terrorism, child sexual exploitation, fraud and scams, and no repeated violations of our other policies.

admitted, contributors may write their own Notes and rate those submitted by others.

Contributors may submit Notes on most public content across Facebook, Instagram, and Threads. This includes content posted by politicians and nearly all other types of organic content, as well as content posted by Meta or Meta executives.[2]

Notes are currently limited to 500 characters and must include a link supporting the context shared in the Note.[3] To ensure that contributors rate Notes based on their content and helpfulness, rather than the identity of who wrote them, author names are not displayed. This anonymity encourages participation by mitigating peer pressure or fears of harassment and prevents ratings from being influenced by the author's identity or how others are voting.

Community Notes is currently limited to the United States. This means that Notes only apply to US-created content, Notes are only visible to US-based users, and only US-based users can become contributors. The product presently functions in six languages: English, Spanish, Chinese, Vietnamese, French, and Portuguese. Language coverage will expand as the product rolls out globally.

B. How the Rating System and Community Notes Algorithm Work

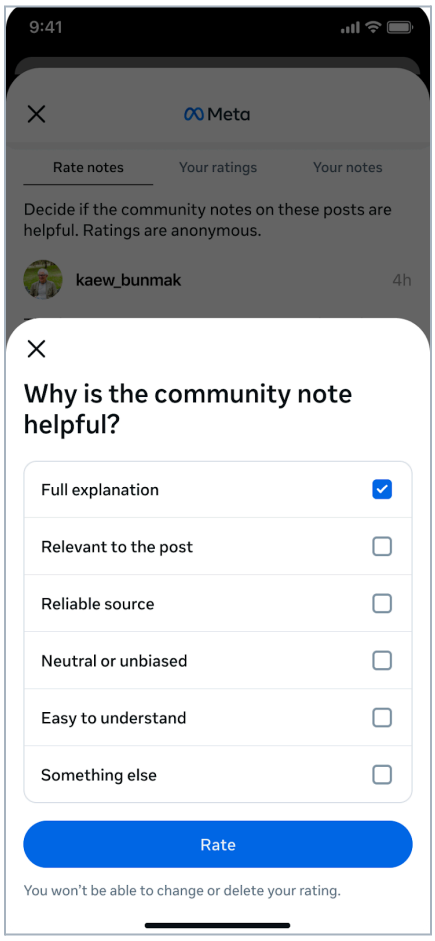
We initially built our Community Notes system using X's open source algorithm, which considers each contributor's rating history. Under this system, Community Notes only appear on posts if (1) others in the community agree they're helpful and (2) contributors who have typically disagreed on a Note's helpfulness in the past now agree that the Note is helpful.

Meta's rating system employs a consensus algorithm that uses separate measures of "helpfulness" and "consensus" to calculate an overall "Helpful Consensus" score. "Helpfulness" is defined as the number of 'helpful' ratings received from contributors who have rated a minimum number of Notes, a requirement for the algorithm to consider rating history.[4]

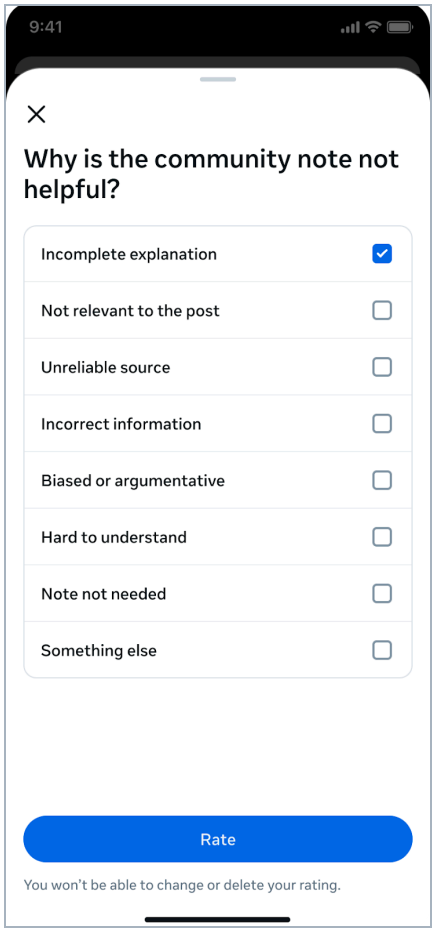
---

[2] At this time, Community Notes cannot be added to advertisements and content with limited audiences such as private posts, messages, Events, Dating, and Marketplace content.
[3] Contributors are required to add a link to a proposed Note, and if the Note is published, the link is shown alongside the Note. Contributors can only add one link to the Note. The links direct users to support for the assertion being made in the Note, such as a link to a newspaper article or a website.
[4] How many users it takes to reach consensus varies depending on the people rating Notes and how many people who usually disagree with each other agree a Note is helpful.

Contributors can rate a Note as “helpful” or “not helpful.” If a contributor rates the Note as “helpful,” they will be asked to explain their reasoning as shown in the following screenshot:



If a contributor rates a Note as “not helpful,” they will be asked to provide feedback by selecting from the following options:



All of these components serve as signals to track agreement and disagreement among contributors that inform our rating system.

"Consensus" is defined as whether those "helpful" ratings come from contributors who previously disagreed, based on how they rated other Notes. To help ensure a diversity of perspectives, the algorithm also considers whether "helpful" ratings come from contributors who likely have differing viewpoints on similar topics based on how they have rated previous Notes.

If the combined "Helpful Consensus" score on a Note exceeds a certain threshold–and the Note does not otherwise violate our Community Standards–the Note will be published. Otherwise, the Note will not be published.

Let's consider the following example of how the algorithm works.[5] Imagine Alex, Harshita, and Maria are Community Notes contributors who rated the same five Notes about football videos. Alex and Harshita rate all 5 Notes as "helpful," while Maria rates all 5 Notes as "not helpful."

Note	Alex	Harshita	Maria
One	Helpful	Helpful	Not Helpful
Two	Helpful	Helpful	Not Helpful
Three	Helpful	Helpful	Not Helpful
Four	Helpful	Helpful	Not Helpful
Five	Helpful	Helpful	Not Helpful

Notes that don't reach broad agreement

Contributors who rate notes similarly won't help a note reach broad agreement, even if they outnumber "not helpful" ratings.

These ratings suggest that Alex and Harshita likely share similar views about football, while Maria holds a different view. However, even with Alex and Harshita's agreement, the Notes will not be published because Alex and Harshita have not disagreed on the helpfulness of Notes in the past.

Later, Alex, Harshita, and Maria rate a new Note adding context to a video about football, with all three finding it "helpful." This agreement may indicate that this Note will be helpful to a broader array of people.

[5] In practice, the algorithm is more complex. This is why 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 to have differing viewpoints. The technical complexity of these types of consensus algorithms is outlined in X's research paper.

Note	Alex	Harshita	Maria
One	Helpful	Helpful	Not Helpful
Two	Helpful	Helpful	Not Helpful
Three	Helpful	Helpful	Not Helpful
Four	Helpful	Helpful	Not Helpful
Five	Helpful	Helpful	Not Helpful
Six	Helpful	Helpful	Helpful

Note reaches broad agreement and is published

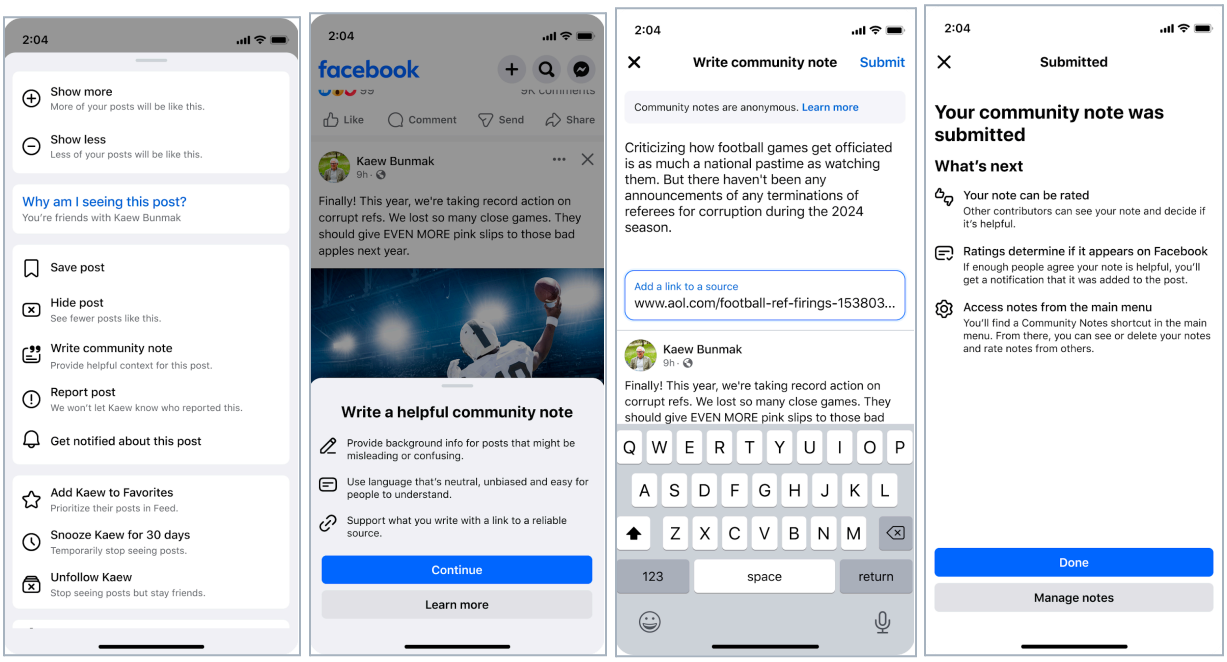
This example shows how Community Notes works. If enough contributors with historically different views agree that a Note is helpful, there is a higher chance the Note will be published.

C. The Community Notes Writing and Viewing Experience

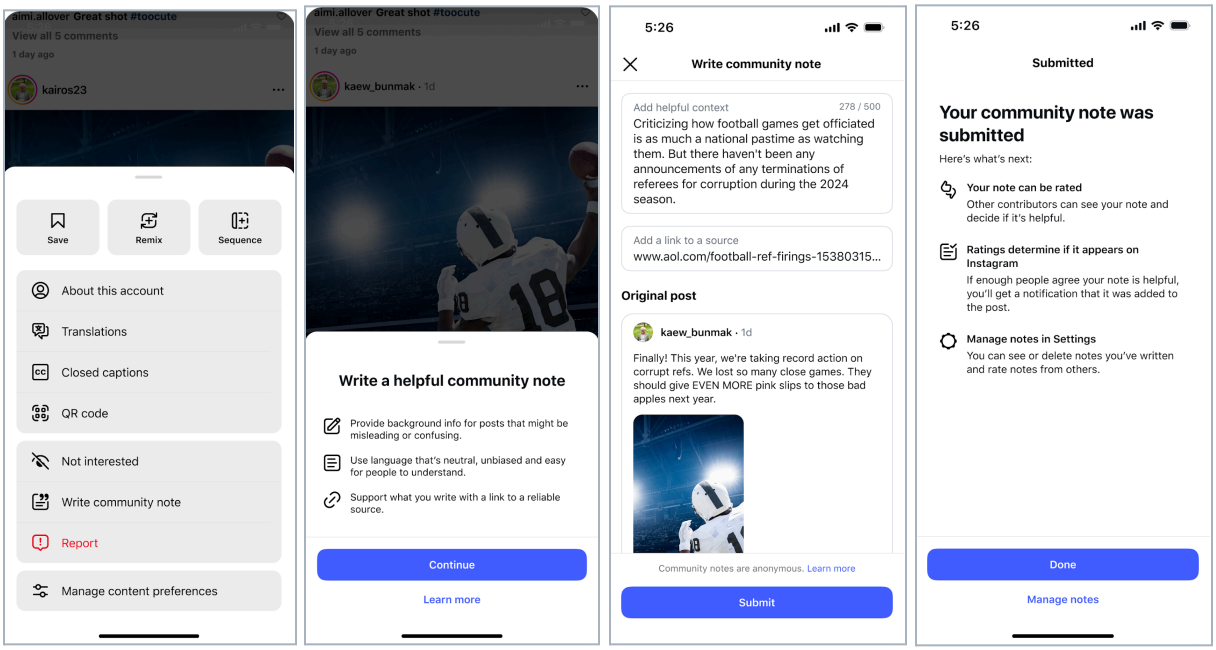
Community Notes contributors have access to a dedicated feed of posts that users tell us may benefit from a Community Note. Contributors may write a Note about a post identified by the community or choose one on their own.

If a contributor sees a piece of content they believe merits a Note, they can click the ‘three dot’ menu next to the post, which opens a menu of options. If the contributor chooses the “Write community note” option, they are taken to a page where they can write their proposed Note. In writing their Note, the user will be prompted to add a link to a source to support their Note. Once completed, the user can press the “Submit” button at the bottom of the screen, where they are taken to a confirmation screen. Example screenshots of the writing flow on Facebook and Instagram are shown below.

Facebook:



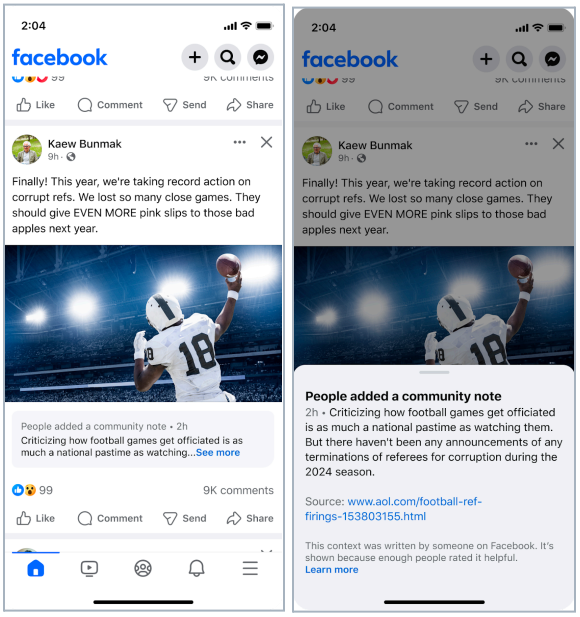
Instagram:



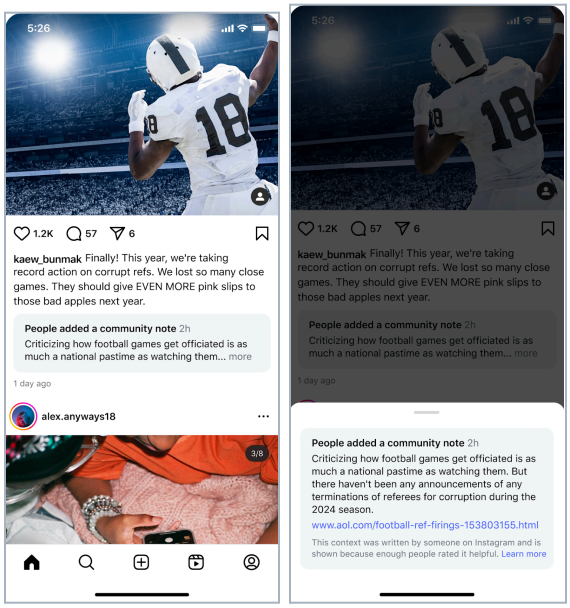
When a Note is published, an abbreviated version of the Note appears as a small banner on the bottom of the underlying post. If a user clicks “see more,” the full Note will pop up over the post. Other users can then read and click on the included link to

see the Note's source.[6] Examples of what Notes look like across Facebook and Instagram are shown below:

Facebook:



Instagram:



[6] Notes are visible even when the underlying post is shared. This includes content shared in private groups.

D. Contributor Retention and Support

We are working on ways to help keep contributors engaged in Community Notes and improve Note quality. New contributors receive an in-product tutorial on how to write and rate Notes, and can access additional resources on Community Notes in Meta's Transparency Center and a dedicated product website. Active contributors can also join a community forum, hosted on Discord, to receive updates on new features and to provide feedback to inform product development. These resources may evolve as the product expands internationally.

E. Safeguards Against Coordinated Abuse

To prevent coordinated manipulation and other abuses, we designed the product with several safeguards. First, because Notes are published only when people having different views find a Note helpful, it is less likely that users can 'game' the system by rating Notes as helpful en masse. Second, all contributors must meet the eligibility requirements described above (e.g., possessing a unique phone number and minimum account age), which help deter the creation of fake accounts. Third, by gradually and randomly admitting users from the waitlist, we help protect against coordinated efforts to add Notes to particular pieces of content or to rate Notes in a particular way. Fourth, we built other additional safeguards into the system, such as spam rate limits and the blocking of high severity violating URL links.

F. Community Notes Are Subject to Our Community Standards

All Community Notes are subject to our Community Standards, and contributors can report Notes for review against those policies both prior to and after publication. This includes removing misinformation that may directly contribute to a risk of imminent physical harm or voter interference.[7]

Because the Community Notes program is designed to be informative, not punitive, content receiving a Note will not experience reduced visibility or reach, nor will users receive strikes. This extends to content posted during periods of crisis or other extraordinary circumstances. Meta will also not alter Note text and will only remove Notes violating our Community Standards.

[7] Community Notes will not replace our current approach to our Misinformation & Harm Policy which, as we state in our Community Standards, involves "remov[ing] misinformation or unverifiable rumors that expert partners have determined are likely to directly contribute to a risk of imminent violence or physical harm to people." We will still rely on expert partners to help make these determinations and continue to remove content that violates our policies.

Currently, Community Notes are not appealable. However, we are exploring adding the ability to request additional ratings from contributors as a potential remediation path for users whose posts receive Notes. If a user deletes a post with a Note on it, the Note will no longer be visible. Similarly, if a post with a Note is made private, the Note will no longer be visible. However, if the post is made public again, both the post and the Note will be visible again.

III. Factors for Omitting Certain Countries from Community Notes

Following the launch of Community Notes in the United States, we are now focusing on our global rollout strategy. As explained in the Introduction, we are seeking the Board's guidance on the country-level factors we should consider when deciding whether any country should be omitted from the global roll out. Given the early stage of product development and limited data from the US beta rollout, our primary interest lies in establishing fundamental guiding principles for these decisions.

Some relevant, non-exhaustive, factors may include[8]:

- ***Low Levels of Freedom of Expression*.** The lack of freedom of expression in a country may influence whether contributors feel comfortable writing and rating Notes without fear of retribution or harassment.

- ***The Absence of Freedom of Press.*** The lack of a free press in a country may affect contributors' ability to participate fully in Community Notes, as there may be limited sources supporting the context they are sharing.

- ***Government Restrictions on the Internet*.** Closely related to freedom of expression, 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).

- ***Low Levels of Digital Literacy.*** Digital literacy – or the ability to use information and communication technologies to find, evaluate, create, and communicate information – may be so low in a country that it should weigh against including that country in our global expansion.

- ***Ability to Achieve Historic Disagreement Required for Consensus:*** The Community Notes rating system relies on agreement among people who

[8] This is merely an illustrative list of possible guiding factors to consider, and is not intended to constrain the Board from considering other factors which may be relevant.

normally disagree. In some countries, there may be less disagreement, or disagreement may manifest differently, potentially resulting in fewer Notes. We are considering how the product may work in contexts where disagreement patterns differ significantly from the United States, such as countries where disagreement is driven more by socio-ethnic divides than political ideology.

We understand that the decision to roll out Community Notes in a particular country will likely depend on multiple factors, none of which may be solely dispositive. These factors will need to be weighed against each other in terms of their relative importance. Therefore, we request the Board's assistance in not only identifying the specific factors to consider for omitting a country from Community Notes, but also in recommending a scalable framework for how to weigh these factors. This framework could include treating some factors as more important than others depending on the particular aspects of a given country.

IV. Conclusion

Community Notes allows more people with more perspectives to add context to more types of posts on our platforms. We value the Oversight Board's insights in helping Meta identify the factors to consider when deciding whether there are any countries to exclude from the global roll out of Community Notes, and how to weigh and assess these factors.