Submission to Meta Oversight Board: Policy Advisory Opinion on Community Notes Expansion[1]

Submitted by: Dr Sanjana Hattotuwa
Date: 7 December 2025

This submission addresses the Board's specific questions whilst examining Meta's broader failure to conduct human rights due diligence, the technical limitations that make community notes unsuitable as fact-checking replacements, and the asymmetric harms Global South communities will face from this transition.

High-level observations

Meta's request asks which countries to exclude from community notes rather than demonstrating why any country should be included. This inverts human rights due diligence. The burden of proof lies with the company to show expansion will not cause harm, not with affected communities to prove they deserve protection.

The Oversight Board faces a choice: provide cover for Meta's predetermined expansion or demand the company meet its human rights obligations before proceeding. The evidence supports only one conclusion: Meta should restore fact-checking partnerships, conduct comprehensive impact assessments with meaningful community consultation, and suspend community notes expansion until independent evaluation confirms the approach functions without enabling the harms the company's platforms have repeatedly caused. **The question is not how Meta should expand community notes but whether expansion should occur at all given the company's failure to establish basic prerequisites: demonstrated effectiveness, appropriate safeguards, adequate language capacity, and credible commitment to human rights principles.**

Community notes could provide value when integrated alongside, rather than replacing, professional fact-checking. Effective implementation would require Meta to restore fact-checking partnerships, use professional assessments to reduce distribution of false content, and allow community notes to provide additional context on borderline cases. This layered approach addresses different aspects of information disorder: professional fact-checkers establish ground truth, algorithmic interventions reduce viral spread, and community contributions surface local context. The current framing eliminates this potential. Meta presents community notes as cost-saving measures replacing labour-intensive partnerships rather than as complementary tools

[1] https://www.oversightboard.com/pc/assessing-metas-plans-to-expand-community-notes/

enhancing platform integrity. Internal research Meta commissioned in 2023 found fact-checking partnerships provided substantially more value per dollar spent than algorithmic interventions alone, findings the company has not publicly disclosed.

Meta's human rights responsibilities regarding product expansion and deprecation

The 2021 Corporate Human Rights Policy established binding commitments

Meta's 2021 Corporate Human Rights Policy made explicit pledges the company is now violating. The policy committed to paying “particular attention to the rights and needs of users from groups or populations that may be at heightened risk of becoming vulnerable or marginalised” and to “identifying relevant such groups for each context, undertaking meaningful engagement to hear their hopes and concerns.” The January 2025 policy changes were announced without consulting civil rights advisory committees, the Oversight Board, or affected communities. The NAACP Legal Defense Fund withdrew from Meta's civil rights advisory committee, stating: “Meta made these content moderation policy changes without consulting or warning us, and many of the changes directly conflict with guidance from LDF and partners.”

On human rights defenders specifically, the 2021 policy stated: “We condemn all threats, acts of intimidation and retaliation, persecution, and physical and legal attacks against human rights defenders… We strive to support their important work.” The revised Hateful Conduct policies now explicitly permit calling LGBTQ+ people “mentally ill” and allow referring to immigrants as “filthy,” directly enabling attacks against categories of human rights defenders Meta pledged to protect. The 2021 policy committed to human rights impact assessments: “When faced with projects or decisions that may significantly affect human rights, we will undertake a human rights impact assessment.” Meta has not disclosed conducting such assessments before eliminating fact-checking or revising hate speech policies. The Oversight Board's April 2025 recommendations noted policies were “announced hastily, in a departure from regular procedure, with no public information shared as to what, if any, prior human rights due diligence the company performed.”

United Nations Guiding Principles require ongoing due diligence

The UN Guiding Principles on Business and Human Rights, which Meta's 2021 policy claims to follow, require companies to “carry out human rights due diligence” including assessing “actual and potential human rights impacts.” Principle 18 specifies this process should “involve meaningful consultation with potentially affected groups and other relevant stakeholders.” Meta's community notes expansion proceeds without disclosed due diligence. The company describes possessing “limited data from the US beta rollout” whilst requesting guidance on global expansion, an admission that impact

assessment has not occurred. Principle 19 requires prevention and mitigation measures proportionate to severity of impacts. Meta's elimination of fact-checking before establishing whether community notes function at scale inverts this obligation.

The request asks the Board to identify factors for "deciding which countries to omit from community notes" rather than which contexts might safely accommodate them. This framing assumes expansion as default, with exclusions requiring justification. Human rights due diligence requires the opposite: companies must demonstrate expansions will not cause harm before proceeding. Principle 20 requires companies to "track the effectiveness of their response" through "appropriate qualitative and quantitative indicators" and "draw on feedback from both internal and external sources." Meta eliminated fact-checking in January 2025, began testing community notes in March 2025, and now solicits guidance on global expansion in December 2025, insufficient time to assess effectiveness or gather external feedback on outcomes.

Challenges in risk assessment for global rollout in polarised contexts

Meta's proposed factors are insufficient and incorrectly weighted

Meta's list of factors for potential country exclusions includes "low levels of freedom of expression," "absence of freedom of the press," "government restrictions on the internet," "low levels of digital literacy," and "ability to achieve the disagreement required for consensus." These factors misunderstand how community notes fail in precisely these contexts. Countries with low press freedom are where professional fact-checking is most critical. When independent media cannot operate, crowd-sourced moderation defaults to state-aligned narratives. Meta's factor list treats absence of press freedom as reason to exclude community notes rather than reason to maintain professional fact-checking. This inverts the human rights analysis: vulnerable contexts require more protection, not less. Similarly, "government restrictions on the internet" correlate with state manipulation of online discourse. Ethiopia's government implemented internet shutdowns during the Tigray conflict whilst operating sophisticated propaganda networks on Facebook. Community notes in such contexts would amplify rather than counter state disinformation because contributor pools skew toward regime supporters, those with continued access, approved speech patterns and incentive structures favouring alignment. "Low levels of digital literacy" appears as potential exclusion criterion whilst research shows these populations face asymmetric vulnerability to sophisticated disinformation. The logical response is strengthened professional fact-checking, not reliance on crowd-sourced contributions from populations lacking tools to assess source credibility. The factor "ability to achieve the disagreement required for consensus" acknowledges the algorithm's fundamental flaw: it requires polarisation to function. In authoritarian contexts with suppressed opposition or societies with cross-partisan agreement on targeting minorities, the algorithm fails precisely because it cannot identify "disagreement." Yet Meta frames

this as reason to exclude countries from community notes rather than reason to question whether the algorithm itself is appropriate.

Language capacity represents a binding constraint

Meta's factor list omits the most determinative element: language. Community notes require sufficient contributor density in each language to generate consensus ratings. Languages with millions of speakers but insufficient Meta user populations, Sinhala, Burmese, Amharic, Tigrinya, will lack contributor bases regardless of other factors. Languages Meta classifies as “low-resource” face systematic disadvantage: fewer moderators, less sophisticated algorithms, higher error rates and slower response times. Code-switching and multilingual content defeat both automated systems and community notes. Indian social media frequently combines Hindi, English and regional languages within single posts. Ethiopian content mixes Amharic, Oromo and English. Somali diaspora communities write in Somali using Latin rather than Arabic script, rendering automated translation systems ineffective. Community notes require contributors who understand all languages within a post, a rare combination.

Conflict contexts require the opposite approach Meta proposes

Meta asks how to assess risks in “contexts of polarisation, conflict or limited human rights protections” whilst pursuing policies that exacerbate precisely these conditions. Research on platform harms during conflict shows consistent patterns: coordinated networks manipulate engagement metrics, state actors suppress opposition voices, propaganda overwhelms accurate information, and violence follows dehumanising rhetoric. The pattern repeats: civil society warns Meta about escalating hate speech, the company responds inadequately due to language constraints, violence occurs, Meta apologises and promises improvement, then deprioritises the same markets when conflicts subside. Community notes could worsen this cycle by eliminating the professional partnerships that provided Meta's only reliable early warning in conflict contexts. Trusted partner networks in conflict zones include organisations with on-the-ground presence, relationships with affected communities, and expertise in local dynamics. These partners contextualise content that appears innocuous to external reviewers but telegraphs violence to intended audiences. Community notes rely on users who may themselves be conflict participants, lack protective protocols for sensitive cases, and have no accountability beyond Meta's opaque contributor scoring.

Recommendations to the Oversight Board

1. The Oversight Board should recommend Meta suspend community notes expansion until the company conducts and publicly discloses comprehensive human rights impact assessments for each proposed market. These assessments must include meaningful consultation with affected communities, independent evaluation of moderation capacity in local languages, and transparent reporting on professional fact-checking partnerships Meta eliminated.
2. The Board should establish minimum thresholds Meta must meet before expanding community notes to any market: sufficient contributor density in local languages,

demonstrated ability to generate consensus within timescales that prevent viral spread, independent validation that notes address substantial percentages of misleading content, and evidence that consensus mechanisms do not systematically disadvantage minority perspectives.

3. Meta must commit to elevated moderation capacity in conflict contexts rather than reduced oversight. Countries experiencing active armed conflict, recent mass atrocities, or patterns of ethnic violence should receive enhanced professional fact-checking, increased human review and expedited escalation processes, the opposite of what community notes provide. The company should establish clear triggers for suspending community notes in contexts where coordinated manipulation or state capture becomes evident.
4. The Board should require Meta disclose internal research on fact-checking effectiveness, community notes performance data disaggregated by country and language, algorithmic audit results for consensus mechanisms, and comparative assessments of intervention effectiveness across markets. This disclosure should occur before any expansion decisions, subject to independent verification by qualified researchers, and updated quarterly to enable ongoing assessment.
5. Meta's 2021 Corporate Human Rights Policy should be enforced as binding commitment. The company must demonstrate how community notes expansion complies with pledges to protect vulnerable populations, support human rights defenders, conduct impact assessments and implement oversight mechanisms. Where compliance cannot be demonstrated, expansion should not proceed.
6. The framing of this request reveals Meta's approach: the company eliminated proven safeguards, imposed changes contradicting its human rights commitments, ignored warnings from civil society organisations, bypassed the Oversight Board's consultation, and now seeks validation for predetermined expansion. The Board's response should centre accountability rather than enablement.