Input into Meta Oversight Board Advisory Opinion
Assessing Meta's Plans to Expand Community Notes

Submission by

Dr Yohannes Eneyew Ayalew

ERC Postdoctoral Fellow

Three Generations of Digital Human Rights Project

The Hebrew University of Jerusalem

Dr Maria O'Sullivan, Associate Professor

Law as Protection Centre, Deakin Law School

Melbourne, Australia

9 December 2025

Table of contents

Submission author information

We provide this submission in our capacity as legal academics with particular expertise in technology and human rights law. We have set out our brief biographies below.

Dr Yohannes Eneyew Ayalew is a European Research Council (ERC) Postdoctoral Fellow at the Faculty of Law, The Hebrew University of Jerusalem. Dr Ayalew is an inaugural Majority World Initiative (MWI) Scholar at Yale Law School in 2023/24. He holds a PhD in Law at the Faculty of Law, Monash University, Australia and dual Master of Laws (LLM) degrees, one in Public International Law from Addis Ababa University, Ethiopia and the other in International Human Rights Law from the University of Groningen, The Netherlands. He also earned his Bachelor of Laws (LLB) from Wollo University in Ethiopia. His work has been published in leading academic journals at the intersection of law, technology and human rights. His most recent publication analyses a third world critique into content moderation: 'A third-world critique of the international human rights-based approach to content moderation' (2025) *Transnational Legal Theory*. His full biography can be viewed here: Yohannes Eneyew Ayalew, Three Generations of Digital Human Rights Project.

Dr Maria O'Sullivan is an Associate Professor and the Theme Lead of the Technology-based Harms Research Stream of Centre for Law as Protection in the Deakin Law School in Melbourne, Australia. She was previously a Deputy Director of the Castan Centre for Human Rights Law at Monash University in Australia. She holds an LLB/BA from the Australian National University, an LLM in International Human Rights Law from the University of Essex in the UK and a PhD in Law from Monash University. Maria is the author of a number of international and national publications on the subject of human rights, public law and technology. Her most recent publication analyses electoral disinformation: 'The Role of Electoral Commissions in Countering Disinformation Implications for Neutrality and Trust' [2025] Public Law. Her full biography can be viewed here: Maria O'Sullivan Deakin University

Submission scope

This submission covers two regions, i.e., Africa and Australia and examines the promises and challenges of rolling out the Community Notes system in those contexts.

Overview of Submission – Key points

1. The risks and opportunities of crowd-sourced and community notes-style approaches.

We recognise the opportunities posed by community notes-style approaches to content moderation. However, we believe that community notes should be complementary to other established mechanisms such as independent fact-checking. There are risks that reliance on community notes as the sole or main source of factual clarification will lead to significant

threats to electoral integrity, political participation, and conflict management in polarised communities.

2. The suitability and adaptability of consensus or bridging-based algorithms, which are employed in systems like community notes to identify and promote content that appeals across divided audiences, to different political contexts and information environments.

We acknowledge that consensus algorithms may be aimed at identifying and promoting content that appeals to different political contexts and information environments. However, the reference in the Board's call for inputs to 'divided audiences' does not adequately describe the level of polarisation in many countries today on issues such as race, immigration and sexual rights. An example of this is the polarisation and use of social media in the defeat of the Australian Indigenous Constitutional Referendum in 2023. A scholar who undertook analysis of social media during that Referendum observed a 'highly polarised communication environment in the dataset under examination, with little or no meaningful dialogue between the Yes and No campaigns evident in the topic model analysis.'[1] We do not believe community notes are an appropriate mechanism to address mis-disinformation in these contexts.
As the Centre for Tackling Hate quite rightly points out:

> The problem is that for a Community Note to be shown, it requires consensus, and on polarizing issues, that consensus is rarely reached. As a result, Community Notes fail precisely where they are needed most.[2]

Indeed, the Centre for Tackling Hate undertook an analysis of 1,060 posts from accounts that were influential in promoting false or misleading claims that contributed to riots targeting migrants and Muslims in the UK. It found that just one displayed a Community Note.[3]
We also believe that the use of community notes does not account for the systematisation of mis/disinformation by powerful actors through the use of LLM technologies. This means that a reliance on 'bridging algorithms' (which are designed to reduce the chance that a single group can dominate the narrative) will be ineffective to adequately address bias. This has been noted by other commentators, both in the US and elsewhere. For instance, Alex Mahadevan from the US-based Poynter Institute has said that 'Bad actors and troll farms have figured out you can flood the system with new accounts to upvote certain viewpoints and get those notes published'.[4] Research into what has been termed "user-generated warfare" has found that

[1] Timothy Graham, 'Exploring a post-truth referendum: Australia's Voice to Parliament and the management of attention on social media' (2024) *Media International Australia* 1-24.
[2] Centre for Tackling Hate, 'Rated not helpful: How X's Community Notes system falls short on misleading election claims', 30 October 2024, p. 4, at: https://counterhate.com/research/rated-not-helpful-x-community-notes/.
[3] Centre for Tackling Hate, 'Rated not helpful: How X's Community Notes system falls short on misleading election claims', 30 October 2024, p. 9 at: https://counterhate.com/research/rated-not-helpful-x-community-notes/.
[4] Alex Mahadevan, cited in Shubhangi Derhgawen, 'Fact check: Are X's community notes fueling misinformation?' *DW,* 5 August 2025 https://www.dw.com/en/fact-check-are-xs-community-notes-fixing-or-fueling-misinformation/a-73315972.

politically motivated users are already manipulating community guidelines to attack content creators on Instagram and TikTok.

3. Meta's human rights responsibilities regarding the expansion and deprecation of products and programs, particularly those addressing misleading information.

We emphasise that Meta has human rights responsibilities under the UN Guiding Principles on Business and Human Rights (UNGPs) to respect human rights and remedy any adverse impacts that their operations may have created or to which they have contributed.
We submit that, as part of this responsibility, Meta should conduct a prior human rights impact assessment before community-notes are implemented, particularly in countries with poor human rights records.

Whilst much debate relating to content moderation has been focused on freedom of expression, we would emphasise the need to also consider *other human rights*. This is important because:

(a) the amplification of incorrect/misleading information about elections can affect the integrity of free and fair elections;

(b) hate speech can have a chilling effect on freedom of association and may propel violence in the community (which has an impact on the right to personal security);

(c) coordinated disinformation campaigns have impacted the right to health. For instance, these campaigns raised significant issues for the right to health during the COVID-19 pandemic; and

(d) social media can be important for advocating for human rights protections. This is illustrated by its use by Indigenous people in Australia to organise campaigns advocating for the right to self-determination and cultural and other rights.[5] For example, a recent Indigenous-led social-media-driven campaign was #SoSBlakAustralia, which sought to stop the forced closure of Aboriginal communities in Western Australia.[6]

(e) the right to fair working conditions and decent compensation for content moderators, data classifiers and workers at content moderation centres in Africa, Asia and around the globe.[7]

Here we also highlight that the proposal to expand the use of community notes coincides with changes to Meta's Hateful Speech policy. As others have noted, these changes mean LGBTQ+

[5] As Bronwyn Carlson and Ryan Frazer note: 'For Aboriginal and Torres Strait Islander populations, who globally tend to fare worse on many social metrics — income, education, life expectancy, political representation, cultural safety — social media can help facilitate vital networks of support, care and knowledge': Bronwyn Carlson and Ryan Frazer , 'Indigenous voices are speaking loudly on social media but racism endures', *The Conversation,* 5 April 5, 2018, https://theconversation.com/indigenous-voices-are-speaking-loudly-on-social-media-but-racism-endures-94287.

[6] Bronwyn Carlson and Ryan Frazer , 'Indigenous voices are speaking loudly on social media but racism endures', *The Conversation,* 5 April 5, 2018, https://theconversation.com/indigenous-voices-are-speaking-loudly-on-social-media-but-racism-endures-94287.

[7] Billy Perrigo, 'Inside Facebook's African Sweatshop' (14 February 2022) *TIME* <https://time.com/6147458/facebook-africa-content-moderation-employee-treatment/> accessed 8 December 2025.

people can be described as "mentally ill," and women can be referred to as "household objects" without the platform deleting the content.'[8] Due to this and other changes in policy, reliance on community notes as the main source of factual correction is not sufficient.

4. Challenges and best practices in risk assessment, monitoring, and mitigation for the global rollout of social media products, particularly in contexts of polarisation, conflict or limited human rights protections.

Community Notes may face the formidable challenge of achieving consensus across deeply divergent perspectives in an increasingly polarised information ecosystem—a difficulty that becomes even more acute in conflict-affected regions in Africa and beyond.[9]

Ongoing armed conflicts in Sudan,[10] Ethiopia during the Tigray war,[11] and in the Sahel region of West Africa[12] illustrate how platforms such as Meta continue to struggle with removing illegal or hateful content even with the support of fact-checking organisations, let alone in contexts where fact-checking services are reduced or discontinued and third-party trusted-flagger programmes are weakened. For example, on 30 October 2025, a viral video purported to show a Sudanese mother and her three children cowering before RSF troops firing their weapons nearby.[13] The video amassed nearly 13 million views on X (formerly Twitter) along with over 80,000 likes and 41,000 reposts and a further 2.4 million views on Instagram. However, a reverse-image search reveals that the footage was originally posted on TikTok on 12 September 2025, more than a month before the fall of El-Fasher, contradicting the narrative promoted by several accounts that reshared it.[14] This example underscores the challenges of mis/disinformation in high-stakes environments and highlights the limits of crowdsourced moderation tools in contexts marked by conflict, fragmented information, and accelerated virality.

---

[8] See discussion in Freya Jetson, 'LGBTQ+ advocates alarmed by Meta's hateful conduct policy changes' ABC News, 10 January 2025, https://www.abc.net.au/news/2025-01-10/meta-hateful-conduct-policy-changes-alarm-lgbtq-advocates/104800042

[9] Liz Orembo and Nerima Wako, 'Meta Discards Fact-Checking: The Fragile Future of Digital Integrity in Africa' (*Tech Policy Press*, 23 January 2025) <https://techpolicy.press/meta-discards-factchecking-the-fragile-future-of-digital-integrity-in-africa> accessed 8 December 2025.

[10] Matt Ford, 'Fact Check: How Fake Content about the Sudan War Spreads – DW – 11/05/2025' (*dw.com*) <https://www.dw.com/en/fact-check-how-fake-content-about-the-sudan-war-spreads/a-74624467> accessed 8 December 2025.

[11] AFP, 'Ethiopia's Warring Sides Locked in Disinformation Battle' (*France 24*, 22 December 2021) <https://www.france24.com/en/live-news/20211222-ethiopia-s-warring-sides-locked-in-disinformation-battle> accessed 8 December 2025.

[12] Issa Sikiti da Silva, 'AI-Generated Disinformation in West Africa Feeds on High Illiteracy Rates' (*Africa in Fact*, 3 July 2024) <https://africainfact.com/ai-generated-disinformation-in-west-africa-feeds-on-high-illiteracy-rates/> accessed 8 December 2025.

[13] Ford (n 10).

[14] ibid.

5. Research into the efficacy of responses to misleading information beyond content removal, such as fact-checking, labelling, reduced distribution, increased friction, and user-generated context. Additionally, research on avoiding bias in such responses.

In terms of the efficacy of responses to misleading information beyond content removal, we highlight the following:

Efficacy of the Australian Electoral Commission Disinformation Register

One interesting example of state responses to misleading information is provided by the possible role that can be carried out by electoral commissions.

The co-author of this submission, Maria O'Sullivan, has written on the role of the "Disinformation Register" [15] by the Australian Electoral Commission (AEC).[16]

The AEC Disinformation Register, established in 2022, is an online platform hosted on the AEC website which lists prominent pieces of disinformation[17] which the AEC has discovered about the Australian electoral process.[18] The stated aim of the disinformation register is to ensure voters "have access to fact-based information about electoral processes".

It is important to note that the Disinformation Register is for the correction of electoral information only – it does not directly address misinformation or disinformation about the broader political process (for instance, disinformation about the electoral policies of political parties). Thus, it is not aimed at disinformation more generally.

An example of this focus can be seen from the 2022 Australian Federal Election, when the AEC discovered an allegation distributed online claiming that the AEC had outsourced counting for the election using vote-counting software. On the Disinformation Register, the AEC corrected that allegation and provided specific information on how it counts votes - clarifying that the Commission does not use vote counting software or voting machines for the House of Representatives and that "Senate ballot papers are counted by hand, then scanned and manually verified in order to allocate millions of preferences, using software developed in-house at the AEC".[19]

In its commentary on the Disinformation Register, the AEC emphasises that it is "not the arbiter of truth regarding issue or political communication and does not seek to censor debate in any way".[20] However, it does note that:

---

[15] AEC, "Disinformation Register", *https://www.aec.gov.au/media/disinformation-register.htm*.

[16] Maria O'Sullivan, 'The Role of Electoral Commissions in Countering Disinformation: Implications for Neutrality and Trust' [2025] *Public Law*.

[17] The Commission defines misinformation as "false information that is spread due to ignorance, or by error or mistake, without the intent to deceive" and disinformation as "knowingly false information designed to deliberately mislead and influence public opinion or obscure the truth for malicious or deceptive purposes": AEC, "Factsheet – Disinformation", *https://www.aec.gov.au/About_AEC/files/eiat/eiat-disinformation-factsheet.pdf*.

[18] AEC, "Disinformation Register".

[19] AEC, "Disinformation Register – 2022 Election", *https://www.aec.gov.au/media/disinformation-register-2022.htm.*

[20] AEC, AEC, *Referendum Report 2023*, Commonwealth of Australia, 2023, p. 17 *https://www.aec.gov.au/About_AEC/Publications/files/referendum-report-2023.pdf Referendum Report 2023*

> … when it comes to the electoral process, the AEC has a responsibility to ensure voters have access to factual information, so they can fully participate in the Australian democratic process. By listing and correcting electoral mis and disinformation, the AEC is assisting to both de-bunk and pre-bunk false narratives about the electoral process, encouraging voters to stop, consider and assess the reliability of the information they are consuming.[21]

In terms of the efficacy of such mechanisms, we note the following:

(i) Not all countries have such bodies. The existence of a central electoral authority or commission is present in some, but not all, jurisdictions globally.[22]

(ii) Despite the fact that the Disinformation Register and other mechanisms employed by electoral bodies can allow electoral bodies to have a voice online, there are significant constraints (both legal and technical) on the ability of these bodies to regulate mis-disinformation. Indeed, the Australian Electoral Commission told a 2024 Australian Parliament Inquiry that:

> 'The AEC does not possess the legislative tools or internal technical capability to deter, detect or then adequately deal with false AI-generated content concerning the election process – such as content that covers where to vote, how to cast a formal vote and why the electoral process may not be secure or trustworthy'.[23]

Against this context, tech companies such as Meta must play a role in supplementing public sector mechanisms to ensure the limitation of misleading information, particularly in the lead up and during elections.

6. Studies that employ quantitative and/or qualitative research methods can help identify and measure country-level factors that might impact the functioning of social media products across different contexts.

We acknowledge Meta's efforts to identify five global benchmark factors for assessing whether to roll out or roll back the Community Notes system. These include: (1) low levels of freedom of expression; (2) absence of press freedom; (3) government restrictions on internet access; (4) low levels of digital literacy; and (5) The ability, currently and in the past, to achieve the disagreement required for consensus. While these factors are important, we submit that additional context-specific considerations should also inform such assessments. For instance, linguistic and cultural diversity is a critical variable in many African countries, where

[21] AEC, *Referendum Report 2023*, p. 17.

[22] See e.g. Australia, Botswana, Canada, India, Poland, Romania, South Africa, United Kingdom, Zambia, Zanzibar.

[23] Reported in 'AEC warns it doesn't have power to deter AI-generated political misinformation at next election', *The Guardian*, 20 May 2024, https://www.theguardian.com/australia-news/article/2024/may/20/aec-australian-electoral-commission-ai-deepfakes-election .

multilingual content ecosystems can complicate community-based moderation. The co-author of this submission, Yohannes Ayalew, has written extensively on the linguistic blind spots observed in Facebook's content moderation practices across three case studies—most notably during the war in Tigray—between 2019 and 2022.[24] His research demonstrates how Meta's moderation efforts have repeatedly faltered in Ethiopia, a country with more than 80 languages, due in part to severe gaps in linguistic coverage and limited investment in methods to overcome these barriers. If Community Notes were to be deployed in such a context, these linguistic and cultural complexities would pose significant challenges to achieving reliable consensus, undermining the system's ability to function effectively and equitably.

Likewise, in Australia, indigeneity and the distinct communicative practices and vulnerabilities of First Nations communities warrant tailored attention. Integrating such contextual factors would enable a more nuanced, equitable, and effective evaluation of the feasibility of Community Notes across different regions.

Factors in implementing community notes

Against that background, we set out our submission on the factors suggested for implementation of community notes:

Factor 1: Low levels of freedom of expression

Across the globe, states recognise the right to freedom of expression, yet its practical protection varies widely. In some countries, the right is robustly upheld; in others, it is routinely undermined through authoritarian techniques deployed both offline and online. Low levels of freedom of expression, for example, often manifest through repressive national security laws, censorship and state control of the media, a climate of fear and self-censorship and pervasive surveillance.[25] They are further reflected in the absence of legal protections for journalists who speak truth to power and citizens who speak out, as well as digital repression in the form of internet shutdowns[26] and social-media bans.[27]

The African Commission's Special Rapporteur on freedom of expression and opinion has similarly highlighted growing concerns, including the deteriorating safety of journalists[28] and

[24] Yohannes Eneyew Ayalew, 'Linguistic Blind Spots in Platform Content Moderation in Ethiopia: The Case of Facebook' (2025) 11 Modern Languages Open <https://modernlanguagesopen.org/articles/10.3828/mlo.v0i0.468> accessed 8 December 2025.

[25] Willem Gravett, 'Digital Neo-Colonialism : The Chinese Model of Internet Sovereignty in Africa' (2020) 20 African Human Rights Law Journal 125.

[26] Felicia Anthonio and Tony Roberts, *Internet Shutdowns in Africa: Technology, Rights and Power* (Bloomsbury Academic 2025).

[27] BBC, 'Nigeria's Twitter Ban: Government Orders Prosecution of Violators' (6 June 2021) <https://www.bbc.com/news/world-africa-57368535> accessed 8 December 2025.

[28] 'Special Rapporteur on Freedom of Expression and Access to Information in Africa, Presented by Honourable Commissioner Topsy-Sonoo Special Rapporteur on Freedom of Expression and Access to Information in Africa'

the proliferation of cybercrime laws[29] which while partly necessary, are increasingly used to criminalise legitimate expression. In such contexts, a weakened culture of free speech would inevitably constrain users' willingness and ability to participate meaningfully in rating and contributing to Community Notes, as repression in both offline and online spaces chills user engagement and deters open dialogue.

Factor 2: The absence of freedom of the press

Similarly, the absence of press freedom is another critical factor that must be carefully evaluated before rolling out the Community Notes system outside the United States. In principle, press freedom allows media outlets and broadcasters to determine their content and report independently in accordance with their editorial policies. Yet, according to Reporters Without Borders'('RSF') 2025 Index, with the exception of South Africa (27/180), Namibia (28/180), Gabon (41/180), Seychelles (45/180), and Mauritania (50/180), most African countries rank far lower, reflecting significant constraints on press freedom.[30] In many jurisdictions, media ownership is highly concentrated in the hands of private actors closely aligned with those in political power, thereby compromising editorial independence and limiting the diversity of viewpoints.

By contrast, Australia, ranked 29/180 by RSF, relatively benefits from a robust press freedom framework that nurtures a culture of open expression, public participation, fact-checking, and investigative journalism.[31] Such an environment provides fertile ground for meaningful engagement with tools like Community Notes. Conversely, in countries where press freedom is restricted, governments often exercise sweeping control over the media landscape, stifling independent reporting and shaping public narratives. In these contexts, online spaces tend to be highly polarised, and mechanisms like Community Notes could be vulnerable to manipulation by powerful state or non-state actors, while dissenting users and marginalised communities may be silenced or discouraged from participating.[32] This concern is echoed in the African Commission on Human and Peoples' Rights' landmark Resolution 630 (2025), which warns of recent regressions in content moderation and reductions in fact-checking capacity by social media platforms. The Commission notably cautioned that "community notes are susceptible to capture by forces that do not respect human rights," pointing out the real risks

(*African Commission on Human and Peoples' Rights*, 29 November 2025) <https://achpr.au.int/en/intersession-activity-reports/freedom-expression> accessed 8 December 2025 para 24.

[29] 'Special Rapporteur on Freedom of Expression and Access to Information in Africa - 85OS' (*African Commission on Human and Peoples' Rights*, 29 November 2025) <https://achpr.au.int/en/intersession-activity-reports/special-rapporteur-freedom-expression> accessed 8 December 2025 para 37.

[30] Reporters without Borders (RSF), 'Media Independence Undermined by Ownership Consolidation and Pressure from Advertisers' (2025) <https://rsf.org/en/classement/2025/africa> accessed 8 December 2025.

[31] Reporters Without Borders (RSF), 'Australia' (2025) <https://rsf.org/en/country/australia> accessed 8 December 2025.

[32] Yohannes Eneyew Ayalew, 'Why Africa Is Sounding the Alarm on Platforms' Shift in Content Moderation' (*Tech Policy Press*, 13 May 2025) <https://www.techpolicy.press/why-africa-is-sounding-the-alarm-on-platforms-shift-in-content-moderation/> accessed 8 December 2025.

of deploying such systems in environments lacking media pluralism and press freedom safeguards.[33]

Factor 3: Government restrictions on the internet

We would like to highlight that government restrictions on the internet do not only happen in authoritarian regimes or in emergency situations.

First case study: Australia's Social Media ban for children

A pertinent example of government restrictions on the internet is the so-called 'Social Media ban' which applies to persons under the age of 16 in Australia. This came into effect on 10 December 2025. This is relevant to content moderation as platforms must now implement more rigorous age-verification mechanisms to ensure compliance. One foreseeable challenge is that some younger users may circumvent the ban by using virtual private networks ('VPNs') or adopting false age credentials, thereby gaining access to platform features, including participating in Community Notes, while many of their peers remain excluded. Such uneven access risks distorting the epistemic balance of the Community Notes system, amplifying certain voices while silencing others, and raises questions about the fairness, representativeness, and legitimacy of crowdsourced moderation tools under restrictive regulatory environments.

Second case study: Internet shutdowns in Africa

Internet shutdowns have become a major and growing threat to the enjoyment of human rights in the digital ecosystem.[34] They occur across political systems—democracies and authoritarian regimes alike—though disproportionately in the latter. A CIPESA study shows that among 22 African countries that imposed shutdowns between 2014 and 2019, 77% were authoritarian, with the remainder classified as hybrid or semi-authoritarian.[35] Access Now, similarly documented that in 2024 alone, 17 African countries intentionally disrupted internet access.[36] When the internet is deliberately switched off, it creates fertile ground for conspiracy theories, misinformation, and harmful content to circulate unchecked. Under such conditions, rolling out a Community Notes system becomes largely ineffective: most users are offline and unable to participate, while others, particularly those in the diaspora with uninterrupted access, may

---

[33] African Commission on Human and Peoples' Rights, 'Resolution on Developing Guidelines to Assist States Monitor Technology Companies in Respect of Their Duty to Maintain Information Integrity through Independent Fact Checking - ACHPR/Res.630 (LXXXII) 2025' (*African Commission on Human and Peoples' Rights*, 29 November 2025) <https://achpr.au.int/en/adopted-resolutions/achprres630-lxxxii-2025> accessed 8 December 2025.

[34] Felicia Anthonio, 'The Kill Switch: How Internet Shutdowns Threaten Fundamental Human Rights in Africa and Beyond', *ISP Digital Future Whitepapers* (Yale Law School 2022).

[35] See Collaboration on International ICT Policy in East and Southern Africa (CIPESA) report (2019), *Despots and Disruptions: Five Dimensions of Internet shutdowns in Africa*, at 4.

[36] Access Now, 'Emboldened Offenders, Endangered Communities: Internet Shutdowns in 2024' (AccessNow 2025) 6 <https://www.accessnow.org/internet-shutdowns-2024/> accessed 8 December 2025.

dominate or even manipulate the system, skewing the epistemic balance. This structural mismatch highlights why internet shutdowns must be treated as a serious barrier to any community-led fact-checking or consensus-building mechanism.

Factor 4: Low levels of digital literacy

We agree that low levels of digital literacy should be a factor militating against any implementation of community notes in many African countries, where large segments of the population remain offline and, even when connected, lack the digital competencies required to participate effectively and meaningfully in such a system.

However, we would like to direct the Board's attention to the fact that the digital divide is not merely one which applies to developing countries. It can also be important in the developed 'Global North' where digital divides can occur across socio-economic, gender or ethnic groups.

This has been recognised by the United Nations Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, who has acknowledged that disadvantaged groups often face barriers to accessing the Internet.[37] He states:

> '… digital divides also exist along wealth, gender, geographical and social lines within States. Indeed, with wealth being one of the significant factors in determining who can access information communication technologies, Internet access is likely to be concentrated among socioeconomic elites, particularly in countries where Internet penetration is low. In addition, people in rural areas are often confronted with obstacles to Internet access, such as lack of technological availability, slower Internet connection, and/or higher costs. Furthermore, even where Internet connection is available, disadvantaged groups, such as persons with disabilities and persons belonging to minority groups, often face barriers to accessing the Internet in a way that is meaningful, relevant and useful to them in their daily lives'.[38]

An example of the existence of a digital divide in a developed country is Australia. Although Australia has a comparatively high rate of internet connectivity, there is still a 'digital divide' and access to technology does not mean that an individual has the capacity to use online platforms.[39] This has been recognised in the Australian Digital Divide report of 2020[40] which notes that:

---

[37] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Frank La Rue to the Human Rights Council, 16 May 2011, A/HRC/17/27, https://www2.ohchr.org/english/bodies/hrcouncil/docs/17session/A.HRC.17.27_en.pdf.

[38] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Frank La Rue to the Human Rights Council, 16 May 2011, A/HRC/17/27, para 61 https://www2.ohchr.org/english/bodies/hrcouncil/docs/17session/A.HRC.17.27_en.pdf.

[39] Gerard Goggin, 'Internet accessibility and disability policy: lessons for digital inclusion and equality from Australia' (2017) *Internet Policy Review,* https://policyreview.info/articles/analysis/internet-accessibility-and-disability-policy-lessons-digital-inclusion-and.

[40] J Thomas, J Barraket, CK Wilson, I Holcombe-James, J Kennedy, E Rennie, S Ewing, T MacDonald, 'Measuring Australia's Digital Divide' The Australian Digital Inclusion Index, 2020, RMIT and Swinburne

- Although internet infrastructure is available to almost all Australians, more than 2.5 million remain offline.
- Increases in digital inclusion for Indigenous Australians have stalled Indigenous Australians living in urban and regional areas have relatively low digital inclusion and recorded no increase over the past year. In 2020, Indigenous Australians' ADII score remains 55.1 and is 7.9 points below the national average. Affordability is a key issue, driven by a disproportionately high use of mobile-only and prepaid connectivity, which carries higher costs per gigabyte than fixed connections. Some Australians are particularly digitally excluded Sociodemographic groups with ADII scores 10.0 or more points below the national average (63.0) are Australia's most digitally excluded. In 2020, these groups include: mobile-only users (43.7), people in low-income households (43.8), people aged 65+ (49.7), and people who did not complete secondary school (51.0).[41]

These statistics are important as digital literacy issues were considered to be an important factor in the influence of mis-disinformation in the 2023 Indigenous Voice Referendum in Australia. This has been recognised by commentators:

> The role of AI in contributing to the No vote has demonstrated an urgent need to close the digital literacy gap. Digital literacy levels impact individuals' susceptibility to consuming misinformation.[42]

In fact, in contexts where digital literacy is low, both the quality and quantity of community notes become questionable. Those who are tech-savvy and understand how the system works are positioned to make the most of it, whereas those who lack such skills may not only produce lower-quality contributions but also engage far less with the system overall. As such, rolling out a Community Notes system without first meaningfully boosting the digital literacy of social media and internet users' risks being dead on arrival, falling short of its intended objectives and failing to foster informed participation.

Factor 5: The ability, currently and in the past, to achieve the disagreement required for consensus [in the community notes algorithm]

As a matter of fact, the Community Notes rating system relies on a baseline level of cross-cutting disagreement — the idea that people who normally hold opposing views can nonetheless converge on what constitutes accurate or misleading information. In some countries, however, such patterns of disagreement may be less pronounced, or may take on fundamentally different forms, potentially resulting in fewer Notes or skewed outcomes. This

University of Technology, Melbourne, for Telstra, https://digitalinclusionindex.org.au/wp-content/uploads/2020/10/TLS_ADII_Report-2020_WebU.pdf

[41] J Thomas, J Barraket, CK Wilson, I Holcombe-James, J Kennedy, E Rennie, S Ewing, T MacDonald, 'Measuring Australia's Digital Divide' The Australian Digital Inclusion Index, 2020, RMIT and Swinburne University of Technology, Melbourne, for Telstra, https://digitalinclusionindex.org.au/wp-content/uploads/2020/10/TLS_ADII_Report-2020_WebU.pdf , pages 6-7.

[42] Dr Tamika Worrell, 'Digital Distortion: How AI Amplified Misinformation in Australia's Voice Referendum – one year on'. Journal of Global Indigeneity, 14 October 2024 https://www.journalofglobalindigeneity.com/post/2761-digital-distortion-how-ai-amplified-misinformation-in-australia-s-voice-referendum-one-year-on .

challenge is particularly acute in many post-colonial African states, where societal divisions often track ethnic, linguistic, or tribal lines.[43] In such contexts, reaching consensus on historical or political matters can be exceedingly difficult, if not impossible. Take Ethiopia as an illustrative example: the figure celebrated as a hero by one ethnic community may be regarded as a villain by another. Thus, disputes about nation-building, historical memory, and collective identity are deeply entrenched and increasingly polarised.[44] A Community Notes system operating in such an environment may therefore struggle to produce balanced or widely accepted annotations. Instead, the system risks reproducing existing fractures — with some groups coordinating to upvote narratives that favour their perspective, while others reject or downvote notes they perceive as hostile to their identity or history. The impact of this dynamic is significant, rather than fostering shared understandings or enhancing the informational ecosystem, Community Notes may inadvertently amplify social fragmentation, entrench competing "truth regimes," and weaken trust in the platform. Without careful contextual adaptation, the system may fail to deliver its promised corrective function in contexts where epistemic consensus is structurally fragile.

Recommendations

Based on the analysis, we recommend that the concerns outlined in paragraphs above be considered in the Board's Advisory Opinion.

First, while Meta's request for an Advisory Opinion is a welcome step, the Board should explicitly urge the company to adopt transparent human rights impact assessments as part of its due-diligence process for any proposed changes or anticipated risk scenarios. In line with the UN Guiding Principles on Business and Human Rights ('UNGPs'), these assessments should be systematically applied in contexts such as elections, public-health crises, and potential outbreaks of violent conflict, particularly in Africa and Australia.

Second, the Board should undertake feasibility studies on the operation of the Community Notes system across different regions and socio-political contexts before issuing a one-size-fits-all Advisory Opinion. Not all African countries have strong records on press freedom or robust protections for freedom of expression, whereas countries such as South Africa and Namibia — and similarly Australia — generally do. In contexts where these protections are weak, Meta should deploy the Community Notes system cautiously and, where necessary, maintain independent fact-checking services rather than replacing them outright.

Third, while we welcome Meta's efforts to identify five global benchmark factors for assessing whether to roll out or roll back the Community Notes system, namely freedom of expression, press freedom, internet restrictions, media literacy, and disagreement, these factors are framed

[43] Mahmood Mamdani, 'Beyond Settler and Native as Political Identities: Overcoming the Political Legacy of Colonialism' (2001) 43 Comparative Studies in Society and History 651.
[44] Mebratu Kelecha, 'Political and Ideological Legacy of Ethiopia's Contested Nation-Building: A Focus on Contemporary Oromo Politics' (2025) 60 Journal of Asian and African Studies 51.

too narrowly. The Board should recommend that Meta incorporate additional factors that are uniquely salient in various regions. These include linguistic diversity, indigeneity, and the structural and historical challenges faced by many Third World countries.[45]

Finally, as Meta shifts gears from professional content moderation by fact-checkers and trained moderators toward a crowd-sourced model like Community Notes, it must take seriously the concerns of African countries and Third World countries more generally. These societies are already grappling with deep structural challenges and have long been not only linguistically excluded but also epistemically marginalised by social media platforms — even under the standard content-moderation regime.[46] Rolling out Community Notes without addressing these inequities risks turning an already uneven playing field into an uphill battle.

Respectfully submitted,

Dr Yohannes Ayalew &

Prof Maria O'Sullivan

[45] Jake Okechukwu Effoduh, 'Digital Colonialism and the Role of Local Intermediaries: Examining Big Tech's Impact on Data Sovereignty and Human Rights in Africa' (2025) 10 Business and Human Rights Journal 301.
[46] Yohannes Eneyew Ayalew, 'A Third-World Critique of the International Human Rights-Based Approach to Content Moderation' (2025) in press Transnational Legal Theory 1 <https://doi.org/10.1080/20414005.2025.2523184> accessed 8 December 2025.