Mandate of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression

Ref.: OL OTH 105/2025
(Please use this reference in your reply)

3 October 2025

Mr. Zuckerberg,

I have the honour to address you in my capacity as United Nations Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, an independent human rights expert appointed by the UN Human Rights Council, to report and advise on human rights issues.

I am writing to you about the **major changes to content policies and the approach to content moderation announced by Meta on 7 January 2025.** I have analysed these changes in my forthcoming report to the UN General Assembly (A/80/341). While I welcome Meta's stated commitment to freedom of expression articulated in the announcement, I am concerned that some of the recent policy changes do not meet international human rights standards and best practices and appear to contradict some of the commitments made in Meta's Human Rights Policy.

Attached to this letter, I share some observations and recommendations on the lack of information on any due diligence carried out prior to the announcement of the policy changes, the policy related to the enforcement of content, revisions to the hate speech policy (renamed the "hateful conduct policy") and the decision to end Meta's Third-Party Fact-Checking program in the United States and to replace it with Community Notes. I do so in a constructive spirit of engagement to encourage Meta to bring its policies in line with all relevant international standards and ensure its platforms are safe for all.

I would be grateful to hear your response to my concerns and to receive further information on the questions below. I am also available to meet with your senior colleagues to discuss these issues:

1. Please provide information on any human rights due diligence that was undertaken by Meta before instituting the 7 January 2025 policy changes and whether any, human rights impact assessment has been or will be carried out on the implementation of these policy changes.

2. Please indicate what measures Meta has taken to address the specific human rights risks associated with the shift from third-party fact-checking in the US to Community notes, and whether and how the lessons learned from this experience inform a possible expansion of Community Notes to other regions.

3. Please indicate the measures that have been or will be taken to assess the impact that Meta's policy changes may have on women, immigrants, LGBTQ+ individuals, and other vulnerable groups.

4. Please indicate any measures that have been taken to assess the impact of policy changes in high-risk countries, including those defined under the Crisis Policy Protocol, and clarify how human rights considerations have influenced decisions about when proactive rather than reactive enforcement is considered more effective.

5. Please also provide information on the scope and nature of changes in curation policies of political content, as well as on any human rights due diligence that may have been conducted in this regard. If no due diligence has been undertaken, please explain why.

6. Meta's transparency reports for the first two quarters of 2025, indicate a significant reduction in overenforcement since Q4 2024. In addition, the reports indicate that violating content actioned under the hateful conduct policy declined from 5.8m pieces in Q4 2024 to 3 million pieces of content in Q2 2025. Does this indicate that Meta is allowing more harmful content on its platform because it is now allowing content that was previously taken down?

7. Please indicate what mechanisms will be introduced to allow for remediation of harm to users on your platforms, as well as the steps taken to ensure that these are accessible in relevant languages, clear and practical to navigate, provide timely relief, and meaningful and effective in practice.

8. Please provide clarification on the impact of Meta's policies to reduce over-enforcement and to respond to my earlier communications on the issue.

In line with our standard practice, this communication and any response received from you will be made public via the communications reporting website after 48 hours. They will also subsequently be made available in the usual report to be presented by Special Procedures to the Human Rights Council. This communication to you will also be referenced in the report I will soon present to the UN General Assembly.

Yours sincerely,

Irene Khan
Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression

Observations and recommendations of the UN Special Rapporteur on the promotion and protection of freedom of opinion and expression

Meta's policy changes announced on 7 January 2025

I. Background

On 7 January 2025, Meta announced several major changes to its content policies and its approach to content moderation in what was described as a return to its commitment to free expression. They included:

- ending of proactive enforcement of content except for "illegal and high-severity violations, like terrorism, child sexual exploitation, drugs, fraud and scams",
- revisions to the hate speech policy (renamed the "hateful conduct policy") including removal of "restrictions on topics like immigration, gender identity, and gender that are the subject of frequent political discourse and debate" and are "too prone to over-enforcement", and
- ending of Meta's Third-Party Fact-Checking program in the United States and replacing it with Community Notes

Following the policy announcements, Meta's Oversight Board expressed its concern that the changes 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". The Board called on Meta to "live up to its public commitment to uphold the UN Guiding Principles on Business and Human Rights".

The Meta Safety Advisory Council, an independent body comprised of online safety organizations and experts, commented that the policy shift "risks normalising harmful behaviours and undermining years of social progress". In particular, the Advisory Council cited concerns that the rollback of protections risks eroding hard-won safeguards that ensure users feel safe and included in online social environments.

II. Relevant international human rights standards

In 2021, Meta adopted a human rights policy, referencing international human rights instruments including the International Covenant on Civil and Political Rights and the UN Guiding Principles on Business and Human Rights, which I welcomed (see my letter AL OTH 212/2021). Meta made an explicit commitment in its Human Rights Policy that the company "will strive to respect" the guiding principles and established a human rights team. Meta also committed to building tools and strategies to address harmful content in countries at high risk of conflict, including by conducting human rights impact assessments and adjusting policies accordingly.

Under the United Nations Guiding Principles on Business and Human Rights, companies have the responsibility to avoid causing or contributing to adverse human rights impacts and to prevent or mitigate such impacts directly linked to their operations, products or services from their business relationships. They are expected to

conduct due diligence through regular human rights risk and impact assessments, mitigate harmful impacts and provide a remedy to those affected. The guiding principles have been endorsed by the UN Human Rights Council and have become a key tool for the private sector's acceptance and implementation of human rights standards.

Companies are obliged to respect international human rights law in their policies, operations and activities. The right to freedom of opinion and expression is guaranteed under article 19 of the International Covenant on Civil and Political Rights ("ICCPR"). Article 19(1) of the ICCPR protects the right to freedom of opinion. This right is absolute and cannot be lawfully restricted. Article 19(2) of the ICCPR guarantees freedom of expression, defined as the "freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice". Freedom of expression may be restricted pursuant to article 19(3), which requires that restrictions must be clear and precise ("provided by law"), and are necessary to achieve "respect for the rights or reputations of others", "the protection of national security or of public order (*ordre public*)", or of "public health and morals".

Under article 20 (2) of the International Covenant, "any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence" must be prohibited. The Office of the High Commissioner for Human Rights has developed the Rabat Plan of Action (A/HRC/22/17/Add.4, appendix) to provide valuable guidance on the factors that should be considered in assessing the appropriateness of prohibiting advocacy of hatred that constitutes incitement.

It is noteworthy that Meta has taken the Rabat Plan into account in moderation decisions. The 2021 Meta Human Rights Report states that "Our Community Standards and Community Guidelines are highly informed by international human rights standards (including the Rabat principles)." In response to a joint communication by the Special Procedures (OTH 20/2024), Meta stated on 29 August 2024 that it had "adapted the principles of the Rabat Plan of Action into actionable content policy tools".

It is also important to note that underpinning all human rights, including the right to freedom of expression, is the right of all persons to equal treatment and to non-discrimination on grounds of race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status (articles 2(1) and 26 of the ICCPR).

III. Compatibility of Meta's policy changes with international human rights standards

1. Human rights due diligence

I am concerned that Meta has not provided to date *any* information as to whether the company conducted any human rights due diligence prior to or since making the major policy changes on 7 January 2025, despite its professed commitment to such procedures under the UN Guiding Principles.

Meta's Oversight Board has called on the company to identify and address the adverse and varied impacts on human rights that may result from the new policies in

different contexts. It has noted that the assessment should examine whether reducing Meta's reliance on automated detection of policy violations could have uneven consequences globally, especially in countries experiencing crises, such as armed conflicts.

I believe that continued and enhanced attention to human rights impact assessments is particularly important as the policy changes announced by Meta in January 2025 do not affect only individualized moderation policies but have broad and significant effect on a wide range of policies and human rights responsibilities of Meta.

I strongly recommend Meta to carry out the human rights impact assessments on the policy changes announced on January 7th without delay, fully and transparently, and to do so on a regular basis. The findings from the assessments should be published and integrated across relevant company processes, tracked to assess whether they are working, and communicated to stakeholders, including externally.

In conducting human rights due diligence over its content policies and practices, Meta should ensure adequate transparency in reporting the results of diligence to stakeholders, with sufficient clarity to assess the adequacy of responses to human rights risk.

2. Content policies and moderation

Content policies and their implementation, including content moderation and curation algorithms, have a significant impact on freedom of expression and public discourse. They may lead to undue restrictions on expression through over-enforcement or take-downs, or they may fail to address harmful content that threatens individuals in vulnerable situations, particularly women and minorities, and impede their ability to participate safely in online spaces.

Meta has a responsibility to ensure that its content policies and content moderation approaches comply with international human rights standards, including the international standards for restricting content, the requirement to prohibit content that is likely to incite violence, hostility or discrimination on certain protection grounds, and the guidance on hate speech provided by the Rabat Plan of Action.

a) Harmful content

Freedom of expression online entails protecting the principle of non-discrimination and ensuring an enabling environment in which all individuals are empowered to speak and access information equally and safely.

I note with particular concern that numerous changes to what was formerly a 'hate speech' policy, now renamed 'hateful conduct policy', reduce the moderation of harmful speech targeting certain protected groups, namely women, gender-nonconforming individuals, and immigrants. Specifically, the policy changes state that "we're getting rid of a number of restrictions on topics like immigration, gender identity and gender that are the subject of frequent political discourse and debate" and are "too

prone to over-enforcement". These topics are inherent about characteristics of groups of individuals.

Meta also appears to have weakened its approach to dehumanizing speech, the policy now stating that it will "allow allegations of mental illness or abnormality when based on gender or sexual orientation" and reference to "women as household objects or property or objects in general". These changes may lead members of these groups to fear harassment or online violence and resort to self-censorship.

The permissive approach to harmful speech based on gender is highly problematic. Gender-based violence online continues to be a major threat, including triggering offline violence. Exercising freedom of expression safely online is essential for women's political, social and economic empowerment, and public representation. The same issues of discrimination and violence also apply to LGBTQ individuals. Gendered disinformation is a serious obstacle to equality in public participation. It is daily reality for women, particularly young women and girls in the Global South, and also for those from the LGBTQ community.

In my report to the UN General Assembly in 2023 (A/78/288) I recommended that companies should conduct regular human rights and gender due diligence, assessing the role of algorithms and ranking systems in amplifying gender-based disinformation. I note that the Oversight Board also recommended that Meta should conduct human rights impact assessments on "potential adverse effects on Global Majority countries, LGBTQIA+ people, including minors, and immigrants, updating the Board on its progress every six months".

The policy changes also fail to include information on the mechanisms available for individuals from vulnerable communities aggrieved by abusive speech. This is particularly critical given that rollbacks in moderating harmful content can easily result in an increase in harmful speech against these groups, thus more urgently requiring remediation mechanisms. It is important for Meta to ensure that there are safe mechanisms for individuals to report hate speech, harassment and incitement to violence and be able to access effective support and gender-responsive remedy without fear of discrimination and further victimization, and systems to record and publish statistics on these violations.

I recommend that Meta's human rights due diligence include specific gender-based due diligence, particularly with respect to its policy changes to reduce enforcement on topics of gender identity and gender. Such gender diligence should review mechanisms for safely reporting gender-based violence online, provide gender-responsive and non-discriminatory remedies, and record and publish statistics specifically on gender-related violations. These statistics should be tracked and labeled distinctly to differentiate them from other violations under a 'hateful conduct' policy.

b) Over-enforcement of content policies

Meta has asserted that one objective of the policy changes is to reduce "over-enforcement". However, it is unclear what criteria, methodology and standards are being used by Meta to reach that determination.

Selective over-enforcement or over-removal by Meta has been a matter of concern to human rights advocates for some time. For instance, I sent a communication to Meta last year outlining bias and systematic removal of Palestinian and pro-Palestinian human rights voices on its platforms (OTH 20/2024). I also expressed concern about the restriction of individual accounts and content published by users of Facebook, Instagram and WhatsApp (OTH 20/2024; OTH 126/2022; OTH 212/2021). I look forward to a response from Meta on how its policy changes to address "over-enforcement" will address the kinds of concerns I had raised, to which I have yet to receive a response.

The only data available to date on the issue of over-enforcement comes from Meta's integrity report for the second quarter of 2025. It claims a 75% reduction in enforcement mistakes on platforms in the United States compared to the previous quarter, and low prevalence of violating content largely unchanged for most problem areas. The report showed a significant decline in take down of hate speech in the first half of 2025, compared to the previous quarter. It is not clear, however, whether the result is due to reduced over-enforcement or fewer pieces of content being defined as "hateful conduct" or to Meta allowing more problematic content to remain on the platform in line with its new policies. It also remains unclear to what extent these enforcement 'mistakes' apply to Palestinian or pro-Palestinian content on its platforms.

I request Meta to provide clarification on the impact of its policies to reduce over-enforcement and to respond to my earlier communications on the issue.

I have previously expressed concerns regarding failures of social media platforms to apply policies consistently across geographical areas and jurisdictions and with respect to vulnerable groups.

Given the likely impact of the policy changes on individuals from vulnerable communities and high-risk areas, I urge Meta to ensure heightened capacity in crisis and conflict situations including coverage in local languages. I particularly urge Meta to ensure that appeals channels are accessible across minority languages so that all users can raise their concerns.

c) Elimination of fact-checking programs in the US

Authoritative third-party fact-checking has long been regarded by Meta as an important tool to counter misinformation and disinformation. Nevertheless, in January 2025 Meta announced the ending of its third-party fact-checking program in the United States in favour of "Community Notes". While the elimination of fact-checking is currently aimed at the United States, the changes will have global impact due to the reach of Meta's platforms.

While community driven review is beneficial in certain situations, it has critical limitations, for instance in addressing disinformation. The weaknesses of community-driven moderation include its susceptibility to capture (where enough users can tip the scales of moderation to spread agendas or disinformation), inconsistent application of standards, lack of consistent expertise, and the requirements of community consensus which can lead to gridlock, especially in polarised situations. Also, crowdsourced

review does not perform to the level of professional fact-checkers. Community-driven fact-checking can be particularly ineffective or even dangerous in situations of crisis, conflict and repression where the online community itself is polarized, deeply implicated in the manipulation of information or unable to distinguish facts from views, and disinformation is likely to lead to real-world harm.

I recommend that Meta assess the human rights impact of its policy to eliminate Third-Party Fact-Checking and introduce Community Notes in the United States. This should include an independent assessment of the respective strengths and weaknesses of both tools in a consultative and transparent manner and tested in diverse contexts, including polarized and fragile security situations. Such an assessment should precede any elimination of third-party fact-checking programs outside the United States.