September 2023

Meta Update: Israel and Palestine Human Rights Due Diligence

Meta

Table of Contents

Introduction

Meta[1] commissioned Business for Social Responsibility (BSR) to carry out an independent Human Rights Due Diligence (HRDD) Exercise on Israel and Palestine for the period of May 2021. This work was scoped as rapid human rights due diligence rather than a full human rights impact assessment in order to enable swift project launch and implementation. Although planning for rapid due diligence began prior to the Oversight Board's September 2021 recommendation in the Al Jazeera case, the exercise sought to be responsive to the Oversight Board's recommendation to:

> *Engage an independent entity not associated with either side of the Israeli-Palestinian conflict to conduct a thorough examination to determine whether Facebook's content moderation in Arabic and Hebrew, including its use of automation, have been applied without bias. This examination should review not only the treatment of Palestinian or pro-Palestinian content, but also content that incites violence against any potential targets, no matter their nationality, ethnicity, religion or belief, or political opinion. The review should look at content posted by Facebook users located in and outside of Israel and the Palestinian Occupied Territories. The report and its conclusions should be made public.*

This HRDD was conducted by BSR in 2021 and 2022 in line with the United Nations Guiding Principles on Business and Human Rights. BSR is an organization of sustainable business experts that works with a global network of the world's leading companies to build a just and sustainable world. BSR's methodology and findings are published in the Insights and Recommendations Report of the Due Diligence Exercise.

[1] On October 28 2021, Facebook, Inc. changed its name to Meta Platforms, Inc. For consistency, this report uses "Meta" to refer to the company both before and after October 28 2021. References to "Facebook" apply only to the social media platform, not the company as a whole. Further, this response references steps taken, or plans to take steps, by Meta as a company regarding a specific entity. Such a statement is not intended to imply that Meta would, or will, take steps regarding all entities. No statement in this response is intended to create — or should be construed as creating — new obligations (legal or otherwise) regarding the application of a policy or procedure to other products or entities. For example (and in contrast to other Meta technologies), WhatsApp is an end-to-end encrypted messaging and calling application with unique human rights touchpoints. This response's discussion of content moderation and related issues on Facebook and Instagram do not apply to WhatsApp. Unless a policy or commitment is specified as applying to WhatsApp, it does not apply to WhatsApp.

This HRDD status update is part of our broader commitment to meaningful transparency about our human rights due diligence and about our integrity work.[2]

[2] Meta's publication of this response should not be construed as an admission, agreement with, or acceptance of any of the findings, conclusions, opinions, or viewpoints identified by BSR, or the methodology employed to reach such findings, conclusions, opinions, or viewpoints. Likewise, while Meta references steps it has taken, or plans to take, that may correlate to points BSR raised or recommendations it made, these also should not be construed as an admission, agreement with, or acceptance of any findings, conclusions, opinions, or viewpoints.

Meta Recommendations Implementation Update

Overview

The HRDD made 21 prioritized recommendations. In September 2022, Meta committed to implement 10 recommendations and partly implement four recommendations and was assessing the feasibility of another six. Meta declined to take further action in relation to one recommendation. This is an update on the status of our implementation efforts in our 2023 Annual Human Rights Report. It covers our work up to June 30 2023.

For ease of reading, we have categorized our update to implementation of the HRDD recommendations to match the order in which they appear in the HRDD executive summary.

We are categorizing the current status of our responses to the recommendations made in the HRDD as follows:

- **Complete**: We have completed full or partial implementation in line with our commitments in response to BSR's recommendation and will have no further updates on the recommendation in the future.
- **In progress**: We are continuing to make progress on our commitments in response to BSR's recommendation and will have further updates on the recommendation in the future.
- **No further updates**: We will not implement the recommendation or have already addressed the recommendation through an action that we already do and will have no further updates on the recommendation in the future.

As of June 30 2023, we had completed implementation of five recommendations, implementation of 10 recommendations was in progress, and there were six recommendations for which we will no longer provide any further updates. This category includes four recommendations that we had already implemented when we published our September 2022 response, one on which we previously announced in our response that we would not be taking further action, and one that we will not be implementing after assessing its feasibility.

Updates to our responses to recommendations

1. Review whether Meta should create policy measures for content that praises or glorifies violence (including indiscriminate attacks, such as violence that is not targeted at any particular person or group).

   Original commitment: Implementing
   Status: In progress

   While a significant volume of such content is covered under our policies on Violence and Incitement, Bullying and Harassment, Violent and Graphic Content, and Dangerous Organizations and Individuals, we are currently conducting a policy review to assess gaps in our policies that address praise or glorification of violent acts.

   We have engaged with over 40 experts as part of this policy process across North America, Europe, the Middle East and North Africa, Sub-Saharan Africa, Latin America, and Asia-Pacific. This included human rights advocates, academics, and researchers focused on violence, trauma, and victim support. Based on this feedback, we are aiming to update our policies in H1 2024.

   These workstreams will likely address some of the concerns around praise and glorification of violence, including between civilians during a conflict. Any resulting policy updates will be in compliance with our legal obligations in this area.

2. Review whether Meta should limit the Dangerous Organizations and Individuals Policy to “support” or “representation” only.

   Original commitment: Implementing
   Status: Complete

   We have concluded a policy review process of our definition of “praise” in our Dangerous Organizations and Individuals (DOI) Policy following extensive analysis of internal data and external engagement with over 100 academics, civil society actors, and other experts across North America, Europe, the Middle East and North Africa, Sub-Saharan Africa, Latin America, and Asia-Pacific. This included security experts, criminologists, political scientists, international lawyers, academics, freedom of expression advocates, human rights organizations, digital rights organizations, civil society organizations (CSOs) focused on countering online hate and extremism, and CSOs supporting the rehabilitation of (former) terrorists.

This review process has examined our current definitions holistically in light of international human rights standards.

Meta is targeting the rollout of a revised policy in H1 2024. Any resulting policy updates will be in compliance with our legal obligations in this area.

3. Review the practice of designating deceased historical individuals under the DOI Policy and assess feasibility of alternative policy approaches to improve transparency and fairness.

   Original commitment: Implementing
   Status: In progress

   We are reviewing our policy and process for delisting designated Dangerous Organizations and Individuals when they no longer meet our designation criteria. Although it does not specifically deal with designated deceased historical individuals, the delisting process review is necessary before reviewing designations of deceased historical figures and assessing whether they still meet our DOI criteria.

   The delisting process review — both policy development and implementation process — is a complex project that will take place for the remainder of 2023. Our goal is to be able to announce changes to our delisting process, if any, in 2024 at the earliest.

   Any future review process of historical designations will be in compliance with our legal obligations in this area.

4. Tier the designation system and strikes for DOI violations to take into account who the organization or individual is and what the violation is (praise, support, or representation) so that the strike is proportional to the violation.

   Original commitment: Assessing feasibility
   Status: In progress

   As part of our policy review process of our definition of "praise" we have examined how our strike system would be applied in a proportional manner to this policy. This is in line with feedback from the Oversight Board and other human rights experts.

   We are targeting the rollout of changes to this policy and related implications for our strike system for H1 2024. Any resulting policy updates will be in compliance with our legal obligations in this area.

5. Provide users with a more specific and granular policy rationale when strikes are applied. This should not just include the category of the violation, but how a post was violating, so that users can better understand the justification, submit an informed appeal, and be less likely to post violating content in the future.

   Original commitment: Implementing in part
   Status: In progress

   As noted in our 2022 HRDD response, we already provide specific granular reasoning when content is removed and strikes are applied in the vast majority of cases. We provide this information in Support Inbox on Facebook and Support Requests and Account Status on Instagram. In 2022, we updated users' messaging to provide more granularity and specificity for 43 different policy violation areas. Work is still underway to expand this to the limited number of areas where we do not yet provide this specificity.

   Given the scale of our enforcement and the fact that users sometimes violate multiple policies, we are sometimes limited in the specificity we can provide when applying strikes.

   We are aiming to introduce further changes that allow us to provide more specific and granular policy rationale for users who violate certain DOI Policies in H1 2024.

6. Increase transparency about Meta's enforcement actions — such as feature limiting and search limiting — and communicate enforcement actions clearly to users.

   Original commitment: Assessing feasibility
   Status: Complete

   We are constantly evaluating and pursuing work to improve our systems and policies for addressing violating content and their transparency to users.

   On Facebook, we are now connecting Profile, Page, Group, and Recommendation surfaces to the Account Status page so that users have a central place to get a more complete picture of their integrity standing, including potential restrictions that apply to their personal profile or Pages they manage and how to appeal.

   We're launching a new pop-up notification feature on Facebook that will alert users if the content they're about to post may violate some of our Community Standards and give them the option to delete the post.

We're also rolling out pop-ups to inform users after a potential violation so that people understand why we have removed their content. That way we are helping users better understand our policies both before and after a potential violation, which is shown to be more effective at preventing re-offending.

In February 2023, we also launched an update to our penalty strikes system to improve clarity about the rules that apply to people on our platforms. Informed by the results of global testing, we started applying read-only feature limits to Facebook Profiles beginning at the seventh strike, as opposed to the second strike. In connection with the update, we published more details about account restrictions in our Transparency Center, including the number of standard strikes that lead to different feature limits.

These changes simplify our penalty system and improve transparency, ensuring users understand the causes of feature limits and consequences of penalties. At the same time, it will expedite the removal of truly persistent violators, who will reach the threshold to have their accounts disabled more quickly upon subsequent violations.

7. Publish key elements of internal community operations resources that help content moderators interpret and apply Meta's content policies so that users can better understand and abide by the policies, excepting adversarial content.

Original commitment: Assessing feasibility
Status: No further updates

We are committed to continued improvement in our transparency efforts. We regularly update our Transparency Center so people can better understand and abide by the policies.

While we won't be publishing internal community operations resources, we will update our Community Standards as part of the launch of our new definition of "praise" and provide guidance on how we interpret our DOI policies.

8. Determine the required market composition (e.g., headcount, language, location) for standby or rapid response capacities for Hebrew and Arabic markets.

Original commitment: Implementing in part
Status: In progress

We are committed to ensuring correct resource investment to address critical events on a sustainable basis. We review market composition regularly, including incorporating insights from human rights due diligence, with this in mind.

Additionally, we made progress in our technology, including ongoing efforts to improve routing by Arabic dialect, as well as strengthening content moderation by reviewers with diverse cultural and linguistic capabilities.

We will continue to review the size of this workforce relative to business and regional needs as these evolve.

9. Continue establishing mechanisms to better route potentially violating Arabic content by dialect/region.

Original commitment: Assessing feasibility
Status: In progress

Following an extensive internal assessment, which included a review of two years of data, we determined that creating specialized routing for a variety of Arabic dialects across our systems would contribute to a greater precision in our Arabic content moderation for high severity content. We are now developing mechanisms to efficiently route content by Arabic dialect.

This will take time, as this is a complex and novel approach requiring a fundamental change in our systems.

10. Assess whether it is feasible and desirable to create a dialect-specific Arabic classifier, working in partnership with Arab linguists and language model experts.

Original commitment: Implementing
Status: Complete

We have conducted analysis on building a dialect-specific Arabic classifier for detection of any content in that language. Based on these results and input from linguists and language model experts, we will add expanded language identification functionality to our systems that will be able to recognize content in different Arabic dialects.

11. Continue work on having functioning Hebrew classifiers.

Original commitment: Implementing
Status: Complete

We noted in our 2022 HRDD response that we had launched a Hebrew classifier that proactively detects and actions violating Hostile Speech content in Hebrew.

We are committed to updating our classifiers to regularly improve accuracy and performance.

12. Adjust the process that allows staff at outsourced providers to add keywords to blocklists to ensure approval by relevant Facebook FTEs.

Original commitment: Implementing
Status: No further updates

We noted in our 2022 HRDD response that we had implemented processes for keywords to be raised by outsourced providers to internal expert teams for assessment and approval for addition. Since implementing this new process, we are not aware of any subsequent comparable issues or errors.

13. Develop a vetting / oversight / quality control process for new additions to hashtag / keyword blocklists.

Original commitment: Implementing
Status: No further updates

We had already implemented such a process at the time of publication of our response in September 2022. Internal teams supporting specific policy areas are responsible for updating and maintaining keywords within Meta's tooling features. Again, since implementing this new process, we are not aware of any subsequent comparable issues or errors with this process.

14. Continue plans to disclose the number of formal reports received from government entities (including the Israel State Attorney Office (ISAO) in Israel) about content that is not illegal but potentially violates Meta content policies. This should take place either quarterly (as part of the Community Standards Enforcement Report) or every six months (as part of the Content Restrictions Report).

Original commitment: Implementing
Status: In progress

As we shared in our Q1 2023 Quarterly Update to the Oversight Board, we are in the process of developing consistent and reliable systems for gathering metrics on the *number of pieces of content removed under the Community Standards as a result of government requests. The objective is to produce government takedown request metrics in the most efficient manner given ongoing challenges including confidentiality obligations and data logging and taxonomy gaps from internal systems.*

We continue to evaluate approaches to building the necessary internal data logging infrastructure to enable us to publicly report this information across the diversity of request formats in which we receive it, but we expect it to be a complex, long-term project. We will provide an update on the timeline for public reporting of these metrics in a future Oversight Board Quarterly Update and in our next annual Human Rights Report.

15. Assess the review accuracy of the DOI Policy enforcement in Arabic across both internal and outsourced teams and including both machine- and human-based review, and address findings (BSR notes this is an ongoing effort).

Original commitment: Implementing
Status: No further updates

This is something we already do. We have a robust accuracy program in place, including for the Arabic market, to help ensure the decisions made by both automated systems and human reviewers are correct.

As part of this program, we consistently strive to improve the accuracy of the enforcement of our policies, including reviewing decisions made by human and automated review and adjusting our operations accordingly.

This is a continuous process and extends across all markets and policy areas.

16. Develop a mechanism to track the prevalence of content that attacks based on specific protected characteristics (e.g., antisemitic, Islamophobic, homophobic content). This might involve, for example, prompting users to mark relevant hate speech content with tags.

Original commitment: Assessing feasibility
Status: In progress

We continue to explore a mechanism to track this but continue to face challenges measuring prevalence with this level of granularity. We remain committed to continue to track the overall prevalence of hate speech on our platforms and evolving our policies to *meet the needs of our users. We will continue to engage with experts in this area as we assess the feasibility of this recommendation.*

17. Establish a structure, protocol, or team to gauge over- and under-enforcement of content policy in a systematic manner during a crisis.

Original commitment: Implementing
Status: No further updates

This is something we already do. In a crisis, we may deploy our Integrity Product Operations Center model to monitor and respond to content trends and other threats in real time. This includes closely monitoring potential over- and under-enforcement of our Community Standards.

18. Increase the capacity of Meta's special escalation channels via more staff and more resources to enable sufficiently prompt response to escalations from Trusted Partners, governments, and other actors in both normal times and times of crisis.

Original commitment: Implementing in part
Status: In progress

Work will continue in H2 2023 to onboard additional Israeli and Palestinian Trusted Partners to escalation channels.

Our ability to onboard new Trusted Partners depends upon sufficient internal resources being available to handle the expected increase in report volumes coming through our escalation channels.

19. Engage in stakeholder engagement and prepare public transparency statement(s) regarding Meta's understanding of its Foreign Terrorist Organization (FTO) and Specially Designated Global Terrorist (SDGT) obligations.

Original commitment: Implementing in part
Status: Complete

As noted in our 2022 HRDD response, although we rely on legal counsel and relevant sanctions authorities to understand our compliance obligations, we regularly review our policies and explore updates to strike an appropriate balance between voice and safety while complying with our legal obligations, and we carry out broad stakeholder engagement in our policy review/development process.

We have done extensive stakeholder engagement as part of our policy review process of our definitions of "praise," "support," and "representation," including with academics, freedom of expression advocates, human rights organizations, digital rights organizations, security experts, criminologists, political scientists, international lawyers, former politicians, journalists, and civil society organizations.

We also briefed Israeli, Jewish, Palestinian, and other civil society and international human rights organizations about this human rights due diligence.

20. Fund public research into the optimal relationship between legally required counterterrorism obligations and the policies and practices of social media platforms. This would address questions such as how the concept of material support for terrorism should be interpreted in the context of social media and whether governments should establish different regulations or interpretations for social media companies.

Original commitment: No further action
Status: No further updates

We have previously stated that we would not be taking any further action on this recommendation. Legal advice is an important foundation to our DOI Policy. As with other legal advice, we do not direct or fund legal guidance in this area for other companies.

We continue to encourage experts to engage with the sanctions authorities that administer sanctions regulations for further guidance.

21. Separate and apart from existing data and law enforcement policies, develop new methods or policies to enable Meta to store content where Meta is under no legal obligation to preserve but where the content may hold potential use for a rightsholder in future remedy processes.

Original commitment: Assessing feasibility
Status: In progress

We are continuing to assess the feasibility of actions in this area in order to support international investigative and accountability processes.

In assessing the feasibility of this recommendation, we are attempting to balance our users' privacy with the needs of international mechanisms established to ensure accountability for atrocity crimes.