CENTER FOR
DEMOCRACY
& TECHNOLOGY

CDT’s Comments to Meta Oversight Board on “From the River to the Sea” Case Bundle

May 21, 2024

The Center for Democracy & Technology welcomes the opportunity to provide comments on cases 2024-004-FB-UA, 2024-005-FB-UA, 2024-006-FB-UA regarding posts that included the phrase “from the river to the sea.”

Meta maintained posts made in November 2023 that included the phrase “from the river to the sea.” The posts were made after the October 7 terrorist attack on Israel by Hamas and reported by users. Users appealed Meta’s action on the grounds that the posts violated Meta’s policies on Hate Speech, Violence and Incitement, or Dangerous Organizations and Individuals.

I. Context is Necessary to Determine Whether the Phrase “From the River to the Sea” Violates Meta’s Policies Regarding Hate Speech, Violence and Incitement, or Dangerous Organizations and Individuals

The phrase “from the river to the sea,” referring to land between the Mediterranean Sea and the Jordan River, abbreviates the phrase “from the river to the sea, Palestine will be free.” The phrase has a long history but has gained prominence in public discourse since the October 7, 2023 Hamas terrorist attack and Israel’s subsequent military response. The meaning of most turns of phrase is context-specific and should be judged on that basis, especially in circumstances where a diversity of speakers are using the same phrase differently. Accordingly, whether the use of the phrase “from the river to the sea” violates Meta’s policies on Hate Speech, Violence and Incitement, or Dangerous Organizations will depend on the context in which it is used.

Meta’s Hate Speech policy prohibits attacks, including dehumanizing speech, statements of inferiority, expressions of contempt or disgust, cursing, and calls for exclusion or segregation against people on the basis of certain characteristics, including religious affiliation and national origin. The policy addresses attacks on people, and expressly excludes such attacks on concepts and institutions. Meta’s policy further clarifies that “content attacking concepts, institutions, ideas, practices, or beliefs associated with protected characteristics, which are likely to contribute to imminent physical harm, intimidation or discrimination against the people associated with that protected characteristic” may violate the policy on hate speech, but additional context is necessary to make such a determination. Meta’s Violence and Incitement policy prohibits content that “incites or facilitates violence and credible threats to public or personal safety.” Meta’s policies prohibit threats of violence that could lead to death or serious injury, and additionally prohibits threats of low-severity violence directed at persons or groups of people on the basis of protected characteristics. Meta’s policy on Dangerous Organizations and Individuals prohibits “organizations or individuals that proclaim a violent mission or are engaged in violence.” Meta’s policy addresses content and platform participation by “Tier I Organizations” and “Tier II Organizations.” Tier I organizations are entities that engage in serious offline harms, including United States Government-designated Foreign Terrorist Organizations, such as Hamas, and Tier II organizations are those that "engage in violence against state or military actors in an armed conflict but do not intentionally target civilians." Meta's policy expressly permits discussion of the actions of Tier I and Tier II organizations and discussion of the human rights of the members of these entities that does not cross into glorification, material support, or representation of the entities or otherwise violate Meta's policies. Meta also prohibits content that glorifies, supports, or represents events designated as terrorist attacks, hate events, multiple-victim violence, hate crimes, and other similar events. Meta designated the October 7 terrorist attack by Hamas a terrorist attack under this policy, thereby prohibiting content that praised, substantively supported, or represented the October 7 attack or its perpetrators.

In light of these policies, Meta's decision to maintain each of the three contested posts must be judged based on the context in which the phrase "from the river to the sea" was used and whether, in light of that context, the post attacked individuals on the basis of their national origin or religious affiliation, incited violence or credible threats to public safety, or glorified, provided material support, or represented actions of a Tier I or Tier II organization.

II. As these cases illustrate, Meta should work to improve oversight of automated processes, moderator guidance, transparency, and commit to periodic evaluations of relevant policies to strike the right balance between user safety and free expression.

Meta's commitment to upholding human rights, and its outsized role in serving as the platform for the free expression and political organizing, make it especially critical that Meta has policies and procedures in place to adequately assess the context in which phrases like "from the river to the sea" are used, particularly in situations where there is ongoing political and armed conflict.

In view of the potential for offline violence and physical injury, takedowns of posts and other strikes against a user account can be appropriate enforcement action in certain cases. Yet, overbroad enforcement can suppress critical dissent, political advocacy, protests, allyship, and news coverage in which the phrase may feature. Further clarification of Meta's hate speech policies and what constitutes incitement of violence under the company's Dangerous Organizations and Individuals policy will offer both users and moderators guidance on what is and is not permissible on the platform. For example, rather than specify that Meta prohibits glorification of "hate crimes," which vary by jurisdiction, Meta should clarify that its Dangerous Organizations and Individuals policy prohibits glorification of both hate crimes and violence against an individual or group on the basis of their immigration status or protected characteristics as defined in Meta's hate speech policy.

Each of the three posts at issue in this case were initially reviewed by Meta's automated system, after being flagged by users, and were not initially flagged for human review, though the second post eventually was examined by human moderators. The fact that Meta relies so heavily on automated review processes for posts that may require nuanced assessments highlights strongly that automated and human review systems should undergo periodic evaluations and human rights impact assessments to ensure equitable and consistent enforcement of policies. Even a narrow and well-tailored content policy can be insufficient at scale and subject to erroneous interpretations, particularly by machine enforcement in situations where context is critical to determining whether a particular post violates the policy. Human rights groups have previously reported that erroneous and inconsistent application of Meta's content policies has led to systematically burdening the free expression rights of marginalized speakers, specifically Palestinian users.

Audits and assessments of automated tools is particularly important as Meta increases its use of large language models trained on scraped language from the internet to moderate content. These models may have hoovered up speech from different corners of the internet and may be vulnerable to the implicit or explicit biases present online. As a result, evaluating and stress-testing these models is critical to ensure models don't scale implicit biases present in training data and burden the rights of speakers, including in languages other than English. Meta should also make the results of these tests public to the degree possible and should detail if and how it adjusts its processes and policies to respond to the test results.

Meta should also evaluate the processes it uses to determine when and whether a post reviewed by an automated system should be reviewed by human moderators. In this case, automated review tools were relied upon when users flagged posts for violating the terms of service. Automated review tools too may have shortcomings and not understand rapidly changing environments. As a result, Meta should, at a minimum, assess when human reviewers should be in the loop to consider context and exercise discretion around borderline posts and whether there are circumstances where human involvement in review of certain content should increase at least temporarily.

These human reviewers also should be provided guidance especially in instances of conflict. Meta should provide guidance to moderators including examples of contexts in which the use of the phrase rises to the level of violating Meta's policies and and when they do not. Meta should accompany these examples with a set of Known Questions or Implementation Standards that allow moderators to assess the context of the speech they are reviewing, particularly to equip moderators to review speech in quickly changing environments. Meta should also engage the experts at the Oversight Board and civil society organizations who have relevant expertise when developing this framework and other moderator guidance.

Additionally, changes to these policies and enforcement of them should be done transparently. Users should be provided notice when a post of theirs has been removed or suppressed due to a user report. User notices should include whether the post has been reviewed by an automated tool, a human moderator, or both to equip users to seek adequate remedy. As CDT has documented in the past, automated content analysis tools may be more error prone in languages other than English due to the dearth of high quality training resources in these languages.