29 June 2023

The Center for Democracy & Technology and the American Civil Liberties Union welcome the opportunity to provide comments on cases 2023-011-IG-UA, 2023-012-FB-UA, and 2023-013-FB-UA, regarding three users' appeals to restore their posts related to abortion in the United States.

The three posts all involve using the word "kill" in reference to abortion policies or beliefs. The first post uses "kill" to reference an abortion itself, while the other two use "kill" as a reference to punishment proposed in state legislation for people who seek abortions. A Meta hostile speech classifier flagged all three posts before human moderators reviewed and removed each post under the Violence and Incitement policy. All three users appealed and after an additional 1-2 human reviews, Meta kept the posts down. Upon selection of these cases by the Oversight Board, Meta restored the posts, saying that the posts did not in fact contain threats or incitements to violence, and did not violate Meta's policies.

As the reproductive rights landscape in the US rapidly changes, it is especially important that Meta allows users to engage in robust discussion and access information about abortion on its platforms. Abortion-related speech can be deeply personal and highly political—the type of speech that has typically received the highest protections under international human rights law and the First Amendment.[1] The same is true for other political speech that may well involve the word "kill," including conversations about school shootings, police killing people, and the death penalty.

To remove political speech on the basis of its purported connection to violence, those strong speech protections require that there be a true threat, incitement to violence, or a direct threat of incitement to violence. The threshold test from the Rabat Plan of Action defines incitement as "an imminent risk of discrimination, hostility or violence";[2] the U.S. Supreme Court has defined a true threat as a "serious expression[ ] conveying that a speaker means to commit an act of unlawful violence"[3] and incitement as speech intended and likely to produce imminent violence;[4] and Meta's own Violence and Incitement Policy requires "a genuine risk of physical harm or direct threats to public safety" to justify a removal.[5] The three cases the Board now considers do not meet these high thresholds.

Both of our organizations support reproductive rights and believe it is important that all users can have frank conversations about abortion on Meta's platforms. Patients should be able to speak freely about their experiences trying to obtain abortions, especially as that ability is further constrained across the US. Both people who have chosen to receive abortions and those who have not need to be able to talk openly about their choices in order to build community with others who have faced a similar decision. Both pro- and anti-choice activists exercising their right to freely assemble need to be able to express their opinions to mobilize action and respond to the wave of abortion-related legislation being introduced across the US.[6] The scale and influence of Meta's platforms makes it critically important for the company to protect abortion-related speech.

---

[1] International Covenant on Civil and Political Rights - OHCHR
[2] OHCHR Rabat Plan of Action
[3] Counterman v. Colorado, No. 22-138, 2023 WL 4187751, at *4 (U.S. June 27, 2023)
[4] Brandenburg v. Ohio, 395 U.S. 444, 447 (1969)
[5] Violence and Incitement - Transparency Center
[6] Data After Dobbs: Best Practices for Protecting Reproductive Health Data - CDT

In this comment we explain why Meta should refine the hostile speech classifier and update its guidance to content moderators to ensure that speech around abortion and other political topics that involve the term "kill" but that do not incite violence are not removed. Ensuring that Meta's content policies and practices protect speech about abortion, reproductive health, and other political speech that uses the word "kill" but that does not incite violence will improve users' ability to engage in important discussions, including those about reproductive rights and abortion access, on Meta's services.

Refine Hostile Speech Classifiers

One of Meta's "hostile speech classifiers" first flagged the three posts in this case. A speech classifier is a blunt, automated tool that by its nature cannot take the context, motivation, or impact of a post into account when evaluating whether it violates Meta's policies.[7] We know little about the specific "hostile speech classifier" Meta employed. In the Tigray Communications Affair Bureau opinion (2022-006-FB-MR), the Board wrote that "hostile speech classifiers" are "machine learning tools trained to identify content subject to Hate Speech, Violence and Incitement, and Bullying and Harassment policies."[8] And, in response to the May 2021 Israel and Palestine Human Rights Due Diligence, we know that Meta launched a Hebrew "hostile speech classifier" to help "proactively detect more violating Hebrew content."[9] Meta already employed an Arabic hostile speech classifier.[10] Meta did not publish the Human Rights Due Diligence Report on the conflict in Israel and Palestine, which makes it harder to understand what these classifiers aimed to address or what makes a "hostile speech classifier" different from other classifiers.

The classifier employed here appears to be trained too broadly because it flags posts for removal that (as Meta agreed, once the Oversight Board selected these cases) do not violate Meta's policies. This poses a serious risk to the ability of users to have conversations about abortion, abortion policies, and the personal experiences associated with abortion, and for people to access information related to reproductive health on Meta's platforms. To improve the accuracy of this classifier, Meta should:

Not rely on the term "kill" alone as a trigger. The term "kill" alone is too common a term to ensure the accurate detection of violent speech. The Oversight Board itself has taken several cases that highlight the ambiguity and overbreadth of flagging the word "kill" on its own. The Board reversed Meta's decision to remove a post that quoted "Kill him!" by Soviet poet Konstantin Simonov, which included the lines "kill the fascist... Kill him! Kill him! Kill!";[11] in the Wampum Belt case, the Board reinstated a post that included a picture of Indigenous artwork titled "Kill the Indian/Save the Man;"[12] and the Board voted to reinstate a video clip from Global Punjab TV where the user "claimed the RSS was threatening to kill Sikhs."[13] Outside of existing Oversight Board cases, it is easy to imagine other situations where a user may post speech that uses the word "kill" but does not incite violence when discussing, for example, school shootings, police killing people, and the death penalty.

---

[7] Mixed Messages? The Limits of Automated Social Media Content Analysis - CDT
[8] Tigray Communication Affairs Bureau - 2022-006-FB-MR
[9] Meta's Human Rights Impact in Israel and Palestine During the May 2021 Escalation - Meta
[10] Independent Report on 'Meta's Human Rights Impact in Israel and Palestine' in May 2021 Released - Lawfare
[11] "Russian poem" case - 2022-008-FB-UA
[12] Wampum Belt - 2021-012-FB-UA
[13] Punjabi Concern over the RSS in India - 2021-003-FB-UA

Instead, if Meta plans to continue using "kill" as part of its hostile speech classifiers to detect violative content, it should make the classifier more narrowly targeted to include other words that, combined with "kill", have a stronger connection to threats or incitement to violence.

Exclude common reproductive terms as triggers. The risk of "kill" being overbroad as a term that flags posts for review is especially true in the abortion context given the view of some anti-choice activists that abortion involves "killing" a fetus, that some states are contemplating whether to treat obtaining an abortion as a felony that could be punished by the death penalty, and discussion by pro-choice activists about the risk that pregnant people may die without access to safe reproductive care. Accordingly, Meta's classifiers should be calibrated so as not treat the word "kill" when used in proximity to "abortion" as an automatic trigger. Meta should similarly ensure that words used by some speakers as synonyms for abortion do not trigger the hostile speech classifiers in combination with "kill", e.g. words like "induced miscarriages," "aborticide," and "termination," because they would also encompass too much political speech.

Additionally, Meta should provide more information about how it is training its hostile speech classifiers in response to state legislation restricting access to abortion medication, particularly how it flags speech about abortion medication under the Restricted Goods policies in these states. Again, Meta should ensure that these classifiers are narrowly tailored, and do not lead to the removal of political or educational speech discussing medication abortions.

Conduct frequent reevaluation. As suggested in the Santa Clara Principles, Meta should also routinely evaluate the effectiveness of the hostile speech classifier that flagged these posts and ensure it is not disproportionately or incorrectly flagging abortion-related content.[14] Meta should immediately evaluate whether the classifier is incorrectly flagging proportionally more abortion-related content today than it did prior to June 2022 and adjust the classifier accordingly. Going forward, Meta should conduct assessments of its classifiers quarterly and include a description of any changes it makes to the classifiers in its quarterly transparency report. Documenting these changes will provide greater transparency to users about the action Meta takes against their content and can be a helpful benchmark for small, less well-resourced platforms who may not have the capacity to closely track the evolving reproductive rights discourse.

Additionally, Meta should ensure that the classifiers it uses are trained on a diverse set of examples of sentences featuring terms they associate with "hostile speech" so they are better equipped to parse relevant and current cultural meanings of words and phrases.[15] Even with better training, however, an automated classifier will have limited ability to assess context and determine how a specific term is used. Moderators should play that role when they subsequently review content that a classifier flags. However, as may have been the case here, moderators may err on the side of agreeing with the classifier whenever they find ambiguity in the post. As we discuss below, moderators should be trained to exercise

[14] Santa Clara Principles

[15] This speaks to a broader issue of the languages and perspectives that Meta uses to train its large language models. For a more detailed examination, see Lost in Translation: Large Language Models in Non-English Content Analysis - CDT.

independent judgment; even with improved moderator training, however, there is a risk that overbroad flagging by the classifier will translate into more erroneous moderator-approved removals. Thus, it remains vital for the classifier to be carefully trained and regularly updated to minimize overbroad referrals to moderators and automatic removals.

Provide more granular context about the moderation decision. As we suggested in 2023-001-FB-UA,[16] moderators should have a way to record how they understood a post, which policy it violated, and why to better allow Meta, the Oversight Board, and at least in certain circumstances the public, to understand where the breakdown in applying the policy occurred. For example, it would be helpful to know if the human moderators reviewing these cases simply accepted the classifier's recommendation, or if they misinterpreted the policy when conducting their own independent review. That, in turn, would inform the relevant policy recommendation: in the former case, the focus should be on improving the classifier and training moderators to make their own decisions, including taking into account factors that are more difficult if not impossible for blunt tools to consider; in the latter, the Oversight Board's recommendations should include rewriting the policy to make it clear that speech about abortion, reproductive health, and other political speech that does not incite violence (even if it uses the word "kill") is allowed.

Meta developed a more granular classifier for hate speech in response to recommendations from 2020-003-FB-UA (Armenians in Azerbaijan) and 2021-002-FB-UA (Zwarte Piet).[17] According to Meta, this classifier allows Meta to notify the user about what type of hate speech it found in the content[18] (although, as the Oversight Board noted, this is currently only available in English and needs to be expanded to other languages). Meta should follow a similar model for the Violence and Incitement policy.

Improve Guidance to Human Reviewers

This case raises several questions about the guidance that human reviewers receive when evaluating posts about abortion and reproductive rights. We do not know from the case summary what guidance reviewers received in the Known Questions or Implementation Standards about abortion (which are "guidelines provided to content reviewers to help them assess content that might amount to a violation of one of Facebook's Community Standards").[19] We also have no information as to why the reviewers found the content violated the Violence and Incitement Policy, or why Meta ultimately reversed the findings of its moderators once the Oversight Board selected this case.

Presumably, seven different moderators reviewed these posts and only one (who was later overturned) believed one of the posts should remain on the platform. The number of human reviewers involved in this case who, as Meta concedes, made the wrong decision speaks to the likelihood that there is a larger issue in the guidance the reviewers have when making decisions about abortion-related content. This could include the language of the Violence and Incitement Policy itself. And it could also reflect problems in the moderator training.

[16] CDT Comments to Meta Oversight Board on Case 2023-001-FB-UA
[17] Oversight Board upholds Facebook decision - Case 2021-011-FB-UA
[18] Case on depiction of Zwarte Piet - Meta Transparency Center
[19] Swedish Journalist - 2021-016-FB-FBR

Information about moderator training. Meta has a mixed history with speech around abortion. Some abortion rights activists allege Meta has restricted posts about abortion for years by classifying the posts as "sensitive" and decreasing their visibility.[20] The day after the Dobbs decision overturned Roe, Meta designated the abortions rights group Jane's Revenge as a terrorist organization under the Dangerous Organizations and Individuals policy (purportedly because Jane's Revenge advocated vandalism).[21] Meta removed a post from Planned Parenthood sharing information about medication abortion in August 2022.[22] And, while not on the platform, Meta banned employees from speaking about abortion following the Dobbs decision.[23]

Especially given this history, Meta should publicly release information about how it trains its moderators to understand and evaluate political advocacy, including speech about abortion and related government policies. This training should include guidance about common tools or tactics of political advocates, including, for example, how they recruit volunteers and encourage others to attend protests, and how Meta understands its value of "expression." Meta should also release information about any bias training it gives its moderators, including the potential for their decisions to be biased by the classifier that initially flags content for their view.

Provide detailed Known Questions about abortion. In addition, Meta should ensure that the Known Questions and Implementation Standards around abortion explicitly highlight the need to preserve political speech that is not threatening. The Known Questions are detailed guidance about specific topics provided to reviewers that go beyond Meta's public content policies.[24] The Known Questions are meant to give reviewers more specific criteria to help them assess whether a post violates one of Meta's Community Standards.

Without access to them, it is hard to evaluate the efficacy of the Known Questions provided to moderators in these cases. But the Oversight Board should ensure that the Known Questions (or Internal Implementation Standards, a document that plays a similar role) clarify Meta's standards for a true threat of violence or incitement to violence. The threat level could reasonably range from something as specific as naming a target or location or as broad as targeting a specific type of person (i.e. abortion provider), but Meta must set this standard so that moderators apply it consistently across content. Additionally, the Known Questions should explain how moderators should distinguish political speech that describes legislation or a speaker's views from speech that is a user's serious expression of intent to harm another person.

The Known Questions about abortion should also include information about words used as synonyms for abortion, medical terms that are commonly used within the context of abortion, and current cultural trends around reproductive terminology. And the Known Questions should instruct moderators to evaluate the post for potential satire or irony when using violence-laden terms or slurs.

---

[20] Meta Was Restricting Abortion Content All Along - WIRED
[21] Facebook Labels Abortion Rights Vandals as Terrorists Following Roe Reversal - The Intercept
[22] Facebook removed a Planned Parenthood post sharing information about abortion pills - The Verge; see also Social media loses ground on abortion misinformation - Axios, where Meta removed groups discussing abortion pills and pro-life candidates.
[23] Meta Clamps Down on Internal Discussion of Roe v. Wade's Overturning - The New York Times
[24] Swedish Journalist - 2021-016-FB-FBR