**Submission to Policy Advisory Opinion 2023L**
***By Jillian C. York and Virginia Kennedy, Electronic Frontier Foundation***

Introduction

The automated removal of abortion related posts containing the word 'kill' fail to meet the criteria for restricting users' right to freedom of expression. Meta has over-removed abortion related content, hamstringing its user's ability to voice their political beliefs. The use of automated tools for content moderation leads to the over-removal of controversial language.

General Over-Removal of Abortion Related Speech

Following *Dobbs v. Jackson Women's Health Organization* Meta began over-removing abortion related speech on their platforms. Shortly after the Supreme Court's decision, multiple news outlets reported that Facebook and Instagram began systematically removing posts related to abortion. These actions went so far as to prompt Amy Klobuchar and Elizabeth Warren to call on Meta to address concerns surrounding the censorship of abortion related posts.

Posts reading, "DM me if you want to order abortion pills but want them sent to my address instead of yours," and "I will mail abortion pills to any one of you. Just message me," were removed within minutes of being posted. When an Associated Press reporter posted "If you send me your address, I will mail you abortion pills" to corroborate the claims, the post was removed within one minute. When the same reporter posted again with the same language but about guns and marijuana instead of abortion pills, the posts were left up. Even abortion related posts that were factually accurate and fully compliant with Meta's policies were removed. Vice reported that a Facebook post stating "abortion pills can be mailed" was flagged within seconds of it being posted. On the other hand, a post stating, "painkiller pills can be mailed," was left up.

Activists who run Facebook groups have voiced their frustration with Meta's vague policies. The inconsistent removal of abortion related information makes it difficult for users to know what is or is not allowed on the site. Meta's Restricted Goods and Services Policy states that "Attempts to donate or gift pharmaceutical drugs" is prohibited and that "Asks for pharmaceutical drugs except when content discusses the affordability, accessibility or efficacy of pharmaceutical drugs in a medical context." In the wake of reports of Meta unjustifiably removing abortion related speech, Meta's spokesperson, Andy Stone, confirmed that content discussing the affordability and accessibility of prescriptions is allowed and that posts were incorrectly removed. This inconsistency in moderation chills legitimate political speech.

Use of the word kill and the necessity of contextualization in moderation

Abortion isn't the only context in which the word "kill" or other controversial terms may be subject to human or automatic removal due to a lack of context.
In one instance, the Oversight Board overturned a decision by Facebook to remove a post accusing Russian soldiers of acting like Nazis. The post contained quotes, including the lines "kill the fascist…Kill

him! Kill him! Kill!" from the poem "Kill him!" by Soviet poet Konstantin Simonov. The Board found that removing the post, and later applying the warning screen, do not align with Facebook's Community Standards, Meta's values, or its human rights responsibilities. The Board additionally emphasized the importance of context in assessing whether content is urging violence.

In another decision from 2021, the Oversight Board overturned a decision by Facebook to remove a post from an Indigenous North American artist under the company's Hate Speech standard. The post in question contained an artwork entitled "Kill the Indian/Save the Man." In this instance, Meta's automated systems identified the content as potentially violating Facebook's Hate Speech Community Standard, while a human reviewer assessed the content as violating and removed it that same day. Meta concurred with the Board that the removal was an "enforcement error", a failure to take into account the context of the use of the word "kill."

A 2022 report by Janny Leung in *Comparative Law and Language* found that even though Meta does not have an explicit policy which favors literal meaning over intended meaning, both the company's automated systems and human reviewers seem geared toward literal meaning, and that Meta's policies also tend to default toward content removal.

As we've noted previously, members of groups often use words that are widely accepted as slurs to reclaim them. For example, members of the lesbian community use the word "dyke" and "dyke marches" take place during pride in many large cities. Members of these already marginalized communities have found their accounts suspended and their posts removed for using the words they are attempting to reclaim. When flagging controversial terms such as "dyke" or "kill" social media platforms should take the context into account when making content moderation decisions.

The dangers of automated removal for "kill" and other controversial text

Poor content moderation has the potential to impose costs on society as a whole including the deprivation of public information and political speech. Automated tools for content moderation have limitations to their usefulness. These tools do not have the ability to recognize nuance or the context of statements. The lack of transparency only adds to the complexity of the issues. In a letter in criticizing the Global Internet Forum to Counter Terrorism ("GIFCT"), civil society organizations pointed out that it was unclear whether protected speech is being censored or if valuable evidence is being destroyed with their automated content moderation tools.

CDT demonstrated that both algorithmic and human-led content moderation includes some subjective (and thus biased) decisions. Given that detailed criteria for content moderation, including enforcement guidelines related to internal policies, are not disclosed, it's difficult to assess the scale and contours of such bias. Additionally, because algorithms can only be trained on known examples, they are more likely to remove similar kinds of content and can be blind to others. The challenges of enforcement of content in languages other than English further exacerbates these issues. The UN Office of Counter-Terrorism (UN OCT) is even beginning to take notice of the limitations of automated content moderation. In a 2021 report, the UN OCT stated "a machine learning model trained to find content from one terrorist organization may not work for another because of language and stylistic differences in their propaganda."

Meta has a responsibility to respect international human rights, consistent with the UN Guiding Principles on Business and Human Rights. Under international human rights law, restrictions to rights such as freedom of expression (art. 19 ICCPR) and freedom of assembly and association (art. 21 ICCPR) can only be justified if there's a legal basis, a legitimate aim, and if they're necessary and proportionate. However, blanket and automatic removal of content without adequately taking into consideration the context in which the word is used, cannot possibly satisfy the condition of proportionality. Indeed, over-broad efforts to remove content can inadvertently result in the suppression of legitimate content, thereby failing to meet the conditions to restrict freedom of expression, civic engagement and activism under international human rights law.

It is of vital importance that online speech is put into its appropriate context before it is removed from the platform. The Rabat Plan of Action provides guidance for companies seeking to remain in compliance with the UNGP's. Meta should consider (1) the social and political context prevalent at the time the post was uploaded; (2) the user's position or status in the society, specifically the individual's or organization's standing in the context of the audience to whom the post is directed; (3) the intent of the user in relation to their audience; (4) the content of the post; (5) the extent of the post, taking into account the post's reach, its public nature, its magnitude, and size of its audience; and lastly (6) the likelihood, including imminence, of harm to result from the post.