Submission to Policy Advisory Opinion 2023-01

By Marlena Wisniak (European Center for Not-for-Profit Law) & Reema Moussa and Jillian C. York (Electronic Frontier Foundation)

Introduction

The automated removal of words such as 'shaheed' fail to meet the criteria for restricting users' right to freedom of expression. They not only lack necessity and proportionality and operate on shaky legal grounds (if at all), but they also fail to ensure access to remedy and violate Arabic-speaking users' right to non-discrimination.

Even within the Arabic-speaking world, the translation of the Arabic word "shaheed" into English (martyr) has created significant controversy. Many Arabic words, phrases, and ideas are not easily translated into English easily due to specific cultural references imbued within Arabic language and culture. This problem is not unique to Arabic—Farsi, Armenian, and other Mid-East based languages bear similar issues.

Given that understanding the use and meaning of the term "shaheed" is largely context-dependent (similar to "jihad", another term over-moderated by Meta), automated content removal tools that lack the complexity to understand or interpret cross-cultural communication are ill-equipped to address the ultimate question that the Oversight Board is evaluating here: is the term "shaheed" being used by Meta's users to incite violence? Or is it used to protest human rights abuses occurring in the Israel-Palestine conflict and other regional disputes and abuses? Alternatively, is it being used to honor and remember Palestinians who have lost their lives—and is there a difference between using the word to honor those who lost their lives innocently or those who were committing an act of violence, or even terrorism?

These are complex questions to consider, and an automated content moderation tool is simply not equipped to consider them. Even non-Arabic speakers may not be equipped to consider them given the cultural context that informs understanding "shaheed" and other Arabic words that don't translate directly into the English lexicon.

The case of Sheikh Jarrah

In May 2021, protests erupted when several Palestinian families were evicted from their homes in the East Jerusalem neighborhood of Sheikh Jarrah. Many Palestinians and allies took to social media to bring attention to this using the hashtag #SaveSheikhJarrah, in both English and Arabic. Users posting content with this hashtag in either language then reported that their posts were being deleted, accounts suspended (or threatened with suspension) by various social media platforms, including Facebook, Instagram, and Twitter.

In response, AccessNow, as well as other civil society organizations such as EFF, 7amleh, and others called for Meta and Twitter to reinstate deleted accounts, "provide transparency on the decision-making processes involved in content takedowns related to Palestine," provide detailed information on "requests submitted by the Israeli Cyber Unit including numbers of complaints received, content removal, account suspensions and other content restrictions," among other things. The campaign was also supported by numerous Palestinian and other public figures. Neither Meta nor Twitter responded publicly to this call to action.

Even prior to the Sheikh Jarrah protests, "dozens of Tunisian, Syrian and Palestinian activists and journalists" found their Facebook accounts deactivated in their reporting on human rights abuses in their respective regions. In a particularly egregious example of Meta's automated translation tools making mistakes leading to content moderation mishaps and even greater harms, in 2017, a Facebook post by a Palestinian man saying "good morning" was incorrectly translated into "attack them," reportedly leading to his arrest by Israeli police.

Regulation of terrorist content online

While binding laws and legislative proposals to regulate terrorist content online have been proliferating, it's important to note that Meta's language rules are not based on legal requirements. As such, there's no legal requirement in banning particular words like 'shaheed.' Meta does so of its own volition, choosing to make an overly broad interpretation of the law.

Meta and other platforms are rarely transparent about the legal basis of content removal, but it seems like they mostly develop their policies based on U.S. and E.U. law. EFF previously assessed that U.S.-based companies such as Meta, Twitter, and YouTube look to U.S. regulations to underpin their policies. As a result, the extremist groups that receive the most focus are typically those on the U.S. Department of State's list of Foreign Terrorist Organizations. Meta, for example, provides a list to moderators that includes photographs of leaders from groups on that list. But although companies use this list as guidance, they are not legally obligated under U.S. law to remove content that comes from these groups.

In the U.S., "material support law" prohibits U.S. persons and entities from providing financial or in-kind assistance to groups on the State Department's list of foreign terrorist organizations. As noted by EFF, the U.S. government has not (at least publicly) taken the position that allowing a designated foreign terrorist organization to use a free and freely available online platform is tantamount to "providing material support" for such an organization, as is prohibited under the patchwork of U.S. anti-terrorism laws. Although the laws prohibit the offering of "services" to terrorist organizations, the U.S. Supreme Court has limited that to concerted "acts done for the benefit of or at the command of another." In February 2023, the Supreme Court heard oral arguments in two terrorism cases, Gonzalez v. Google and Twitter v. Taamneh. While outcomes of these cases certainly have the potential to severely harm freedom of expression and contribute to holding platforms liable for facilitating terrorist content online, platforms still have immunity for user-generated content at this time. And U.S. courts have consistently rejected efforts to impose civil liability on online platforms when terrorist organizations use them for their communications. The Supreme Court has limited these restrictions to concerted "acts done for the benefit of or at the command of another."

In the EU, the Terrorist Content Regulation requires platforms to act within only one hour upon removal orders issued by authorities, an extremely short deadline. As much as this regulation poses a clear risk of over-removal of legitimate content, it certainly doesn't require platforms to act in that direction. Relatedly, there's no legal requirement in banning particular words like 'shaheed" or words that could be construed as praising terrorism.

A key issue for moderating terrorist content online globally is that there is a lack of universally agreed upon definition of terrorism. In resolution 1566 (2004) on threats to international peace and security caused by terrorist acts, the UN Security Council defines terrorist offenses broadly; as ECNL previously pointed out, the definition of terrorism has been subject to significant debate, with different organizations and national governments operating under different understandings of terrorism. The act of labeling certain groups as "terrorists'' is a normative claim; the line between terrorism and other political violence is not always clear. The Global Internet Forum to Counter Terrorism (GIFCT) warns that, in the context of a "highly politicized context within which counterterrorism takes place have resulted in government overreach." The UN Special Rapporteur on Counter-terrorism and Human Rights, Fionnuala Ní Aoláin, also questioned social media companies' definitions of controversial terms like "terrorism" and "terrorist organizations'' in her 2018 letter to Mark Zuckerberg.

To add onto the lack of clarity and consistency, Meta isn't transparent about how it designates or classifies certain groups, nor how it interprets the law. There's no law that prohibits Meta from sharing this information; quite the opposite, as the EU Terrorist Content Regulation establishes some transparency requirements for

authorities and social media platforms, as well as redress mechanisms for users. EFF previously stressed that Meta's combination of ever-increasing automation and Meta's vague and opaque rules (none of which cite any legal requirements) make it impossible for users in affected countries to understand what they can and cannot say. Meta has a responsibility to be transparent to its users and let them know, in clear and unambiguous terms, exactly what content can be discussed on its platforms. The Oversight Board itself has repeatedly criticized the vagueness of rule creation, interpretation, and enforcement. Meta responded by clarifying the meaning of some of the terms, but left some ambiguity and also increased its unguided discretion in some cases.

Beyond complying with counterterrorism laws, Meta also has a responsibility to respect international human rights, as 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 that praises terrorism—such as 'shaheed'— 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 terrorist 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.

This lack of proportionality is exacerbated as moderation of terrorist content is increasingly automated through algorithmic systems. The Center for Democracy and Technology (CDT) cautioned that while such systems can be helpful in moderating content at scale, they have significant limitations. ECNL further warned that these systems often exacerbate and accelerate existing challenges related to content moderation, not least related to the lack of transparency and understanding of local context. Indeed, algorithmic systems based on keyword detection and language models are not able to fully capture the nuance of statements, particularly when it comes to irony or culturally-specific references. As a coalition of civil society organizations concluded in their joint letter to the Global Internet Forum to Counter Terrorism (GIFCT), this has led to the inadvertent deletion of legitimate speech such as journalism, satire, art, anti-terrorism critique, and documentation of human rights abuses.

Between January and September 2021 alone, Meta removed 25.9 million pieces of content as violating its community standard on terrorist content, according to its own internal reports. Unfortunately, Meta and other platforms do not publicly disclose how many removal decisions were later reversed because of a mistake. Prior errors and overbroad enforcement measures were reported for Chechnya, Kurdish activists, and Al Jazeera. EFF reported that Meta reversed the decision only

after the Oversight Board selected the case, as it did in two other similar cases. In another case, Meta apparently misplaced important policy guidance in implementing the DIO policy for three years.

The indiscriminate removal and use of algorithmic systems for moderating terrorist content such as 'shaheed' also violates users' right to non-discrimination, as these systems disproportionately suppress the speech of users and groups that are already marginalized and vulnerable. CDT showed 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 biased towards removing certain kinds of content and can be blind to others. 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."

As ECNL stated, over-enforcement of policies pertaining to terrorist content or violent organizations has inadvertently led to the removal of legitimate content of Muslim and Arabic-speaking communities, thereby violating their right to non-discrimination in addition to their freedom of expression, assembly and association. Intentional or not, content exposing human rights abuses or criticizing powerful actors can be erroneously flagged as violative, and thus removed. ECNL partners Hayat-Rased and TÜSEV documented such discriminatory enforcement in Jordan and Turkey, respectively. In their human rights impact assessment of GIFCT, Business for Social Responsibility made a similar assessment. Additionally, analysts from the Brennan Center for Justice found that content uploaded by Muslim users is disproportionately policed on major social media platforms, in comparison to content in support of white-supremacist organizations.

Finally, all the above issues are amplified given that sanctions for posting terrorist content online, or praising terrorist acts, are particularly severe. They often lead to deplatforming, where the user's account is permanently suspended, preventing them from creating a new account. If not banned, affected users are at the very least sanctioned through heavy strikes. Another amplifying factor is that major social media platforms, including Meta, use the GIFCT hashtag database. While this can help smaller platforms detect previously identified terrorist content, it also means that a single database may be used broadly across the Internet, and errors are thus multiplied.

Recommendations

It is our opinion that option 2—the removal of content using "shaheed" to refer to individuals designated as dangerous under Meta's policies only where there is praise, support, or a signal of violence—is the most salient option. Meta has noted that the option better aligns with both the principles of international law and their "value of voice" but "could be perceived as promoting voice over the value of safety." We believe that this maximization of voice is an imperative for ensuring the free expression of Arabic-speaking users and allowing people to use "shaheed" in the manner appropriate to their cultural context.

Specifically, with respect to automated moderation, we feel that it is vital that Meta does not automatically remove the term "shaheed" in any instance, but rather only employ automation in limited circumstances, in order to flag content which falls in the aforementioned category for human review.

Furthermore, in alignment with recommendations from the majority of civil society organizations, we stress the importance of Meta's continued consultation with civil society, particularly groups with relevant regional, cultural, and Arabic-language expertise.