DigitalRightsFoundation
"KNOW YOUR RIGHTS"

DIGITAL RIGHTS FOUNDATION PUBLIC COMMENT SUBMISSION ON OVERSIGHT BOARD CASE REGARDING THE IRAN PROTEST SLOGAN

Submission Authors: Shmyla Khan and Noor Waheed
Submission Date: 18th October, 2022

Digital Rights Foundation Pakistan (DRF) welcomes the opportunity to provide comments on case 2022-013-FB-UA , regarding the Iran Protest Slogan.

It is imperative for Meta to maintain a balance between preventing violence against any figure, particularly political and religious, and the protection of its users' right to protest, and their right to political speech. Meta's 'Violence and Incitement Policy' makes allowances for "non-credible" "aspirational or conditional threats aimed at [violent actors]" and distinguishes between non-threatening rhetorical use of violent language with actual "high severity violence", however no detailed standard exists to make these determinations. Overall, it must be recognized that legitimate expressions of anger and dissatisfaction against powerful actors should not be policed and can rely on language of violence.

To ensure consistency in its application for newsworthiness allowances to the right of citizens to protest, Meta should establish a threshold for what constitutes "high severity violence". Additionally, determinations of whether calls for violence against a political leader constitute a "credible threat" should be in light of cultural context. For example, in the context of Iran, "Death to Khameni" is considered the equivalent of "Fuck, Trump" (Source: "*Facebook Says 'Death to Khamenei' Posts Are OK for the Next Two Weeks*," Vice). It is a placeholder slogan meant to be critical of the regime rather than a call to action. The issue at hand is not just definitional, rather whether Meta has the cultural context and dedicated resources to make these determinations.

When considering limitations on "calls for violence" against prominent public figures, Meta's content moderation systems should assess whether or not those statements (besides being credible) pose an imminent/immediate threat that could trigger an inflammatory response especially in "high-risk", "volatile" geo-political situations. Such considerations should not be left to automated systems ("Iran: Meta must overhaul Persian-language content moderation on

Instagram," Article 19) that have been known to unconsciously perpetuate political bias such as in the case of reporting on Israel and Palestine. Digital Rights Foundation has found in its monitoring of Farsi/Dari and Pashto content on Facebook that harmful content in these languages is rarely flagged given that such determinations require nuanced understanding of the context (policy brief can be shared upon request). Instead any special allowances to allow "Death to" statements on the platform should be overseen by a team of human reviewers who are privy to the geographical, cultural, linguistic and contextual factors at play. Furthermore calls for violence against political figures should not be solely tied to language and keywords such as "Death to"; violence can be fomented through other means such as blasphemy accusations which in some contexts are more dangerous than directly violent statements. Additionally measures to ensure allowances for certain statements in times of "unrest" should be standardized according to international human rights law to ensure such determinations are not politicized and subject to selective application ("Civil Liberties Groups Urge Social Media Platforms to Better Protect Free Flow of Information in Crisis Zones," EFF).

Given the scope and reach of Meta's various platforms and their subsequent ability to trigger unrest or alternatively highlight human rights violations, Meta should consider whether certain violent slogans rhetorical or otherwise should be censored based on a "risk-assessment" or "volatility-assessment" conducted by their dedicated team of human reviewers. Measures and guidelines should be in place to ensure that these human assessments are not reproducing cultural and racist stereotypes that categorize local populations as inherently more prone to violence or perpetuate harmful political biases that disproportionately favor the status quo and can undermine sincere human rights efforts in these countries. The composition of this team of reviewers and the contents of these assessments should be made publically accessible for transparency and accountability. Considerations regarding removal of speech pertaining to public figures should be balanced with other rights that might be impacted by a potential removal such as freedom of assembly and association (UN Human Rights Committee, 1996, para. 12). In the event Meta has to make a decision to take down content that calls for "death" against a major political/religious figure for reasons of social unrest/volatility, the intervention should be narrowly tailored. Users posting the content in a manner that is not in violation of the "high severity violence" requirement or the community guidelines at large should not be penalized, barred from posting or have reach restrictions.

This case also highlights the need for greater transparency from platforms in their content moderation decisions. Since no reasons were provided for the decision in this case, it is unclear whether the automated system took into account the comments under the post indicating credible plans for violence or whether contextual factors were considered. Regardless of whether a decision was taken through a human in the loop or through automated systems, factors taken into account and detailed reasons should be provided to the user for greater transparency and accountability. Secondly, access to appeal mechanisms is one of the most crucial aspects of

speech regulation, which is missing in this case. Users should have recourse to a responsive, speedy and transparent appeals mechanism as a core feature of content moderation as opposed to optional mechanisms subject to availability of resources. Lastly, the facts of the case point towards over-reliance on automated decision making at a systemic level as one of the long-term implications of the COVID-19 pandemic, as it was used as an excuse for lack of recourse to appeal over two years into the pandemic.

*To read the Oversight Board's full decision on this case:
https://www.oversightboard.com/decision/FB-ZT6AJS4X

**To see all submitted Public Comments:
https://oversightboard.com/attachment/3376041612608638/