Iranian Woman Confronted on Street

2023-032-IG-UA

Summary

The Oversight Board has overturned Meta's original decision to take down a video showing a man confronting a woman on the streets of Iran for not wearing the hijab. The post did not violate the Violence and Incitement rules because it contains a figurative statement, rather than literal, and is not a credible threat of violence. Shared during a period of turmoil, escalating repression and violence against people protesting, access to social media in Iran is crucial, with the internet representing the new battleground in the struggle for women's rights. As Instagram is one of the few remaining platforms not to be banned in the country, its role in the anti-regime "Woman, Life, Freedom" movement has been immeasurable, despite the regime's efforts to instill fear and silence women online. The Board concludes that Meta's efforts to ensure respect for freedom of expression and assembly in the context of systematic state repression have been insufficient and it recommends a change to the company's Crisis Policy Protocol.

About the Case

In July 2023, a user posted a video on Instagram in which a man confronts a woman in public for not wearing the hijab. In the video, which is in Persian with English subtitles, the woman responds by saying she is standing up for her rights. An accompanying caption expresses support for the woman and Iranian women standing up to the regime. Part of the caption, which also criticizes the regime, includes a phrase that translates as, "it is not far to make you into pieces," according to Meta.

Iran's criminal code penalized women who appeared in public without a "proper hijab" with imprisonment, a fine or lashes. In September 2023, Iran's regime approved a new Hijab and Chastity Bill under which women could face up to 10 years in prison if they continue to defy the mandatory hijab rules. The caption in this post makes it clear the woman in the video has already been arrested.

First flagged by Meta's automated systems for potentially violating Instagram's Community Guidelines, the post was sent for human review. Although multiple reviewers assessed the content under Meta's Violence and Incitement policy, they did not come to the same conclusion, which, in combination with a technical error, meant the post stayed up. A user then reported the post, which led to an additional round of review, this time by Meta's regional team with language expertise. At this stage, it was determined the post violated the Violence and Incitement policy, and it was removed

from Instagram. The user who posted the content then appealed to the Board. Meta maintained its decision to remove the content was correct until the Board selected this case, at which stage the company reversed its decision, restoring the post.

Key Findings

The Board finds the post did not violate the Violence and Incitement Community Standard because it contains figurative speech, rather than literal, and is not a credible threat of violence that is capable of inciting offline harm. While Meta originally removed the post partly because it assessed the phrase, "it is not far to make you into pieces," as a statement of intent to commit high-severity violence – targeting the man in the video – it should not be interpreted literally. Given the context of widespread protests in Iran, and the caption and video as a whole, the phrase is figurative and expresses anger and dismay at the regime. Linguistic experts consulted by the Board noted a slightly different translation of the phrase ("we will tear you to pieces sometime soon"), explaining that it conveys anger, disappointment and resentment towards the regime. Rather than triggering harm against the regime, the most likely harm that would result from this post would be retaliatory violence by the regime.

While Meta's policy rationale suggests "language" and "context" may be considered when evaluating a "credible threat," Meta's internal guidance to moderators does not enable this in practice. Moderators are instructed to identify specific criteria (a threat and a target) and when those are met, to remove content. The Board previously noted its concern about this misalignment in the Iran Protest Slogan case, in which it recommended that Meta provide nuanced guidance on how to consider context, directing moderators to stop default removals of "rhetorical language" expressing dissent. It remains concerning there is still room for inconsistent enforcement of figurative speech, in contexts such as Iran. Furthermore, as automation accuracy is impacted by the quality of training data provided by humans, it is likely the mistake of removing figurative speech is amplified.

This post was also considered under the Coordinating Harm and Promoting Crime Community Standard because there is a rule prohibiting "content that puts unveiled women at risk by revealing their images without [a] veil against their will or without permission." The policy line has since been edited to prohibit: "Outing [unveiled women]: exposing the identity of a person and putting them at risk of harm." On this, the Board agrees with Meta that the content does not "out" the woman in the video and the risk of harm had abated because her identity was widely known and she had already been arrested. In fact, the post was shared to call attention to her arrest and could help pressurize the authorities to release her.

As Iran is designated an at-risk country under Meta's crisis policies, including the Crisis Policy Protocol, the company is able to apply temporary policy changes ("levers") to address a particular

situation. While the Board recognizes Meta's efforts on Iran, these have been insufficient to ensure respect for people's freedom of expression and assembly in environments of systematic repression.

The Oversight Board's Decision

The Oversight Board has overturned Meta's original decision to take down the post.

The Board recommends that Meta:

- Add a lever to the Crisis Policy Protocol to make clear that figurative (i.e., not literal) statements that are not intended to, and not likely to, incite violence, do not violate the Violence and Incitement policy line that prohibits threats of violence in relevant contexts. This should include developing criteria for at-scale moderators on how to identify such statements in the relevant context.

*Case summaries provide an overview of cases and do not have precedential value.

Full Case Decision

Section 1: Decision Summary

The Oversight Board overturns Meta's original decision to take down a video that shows a man confronting a woman on the streets of Iran for not wearing the hijab. Meta removed the post from Instagram because of the following line in the caption – "it is not far to make you into pieces" – which the company read as a threat, targeting the man in the video who approached the woman for not wearing a hijab. The Board finds the post did not violate the Violence and Incitement policy because the relevant statement is figurative, rather than literal, and given the context, does not constitute a credible threat of violence. After the Board selected the case, Meta initially upheld its decision to remove the post – however, before submitting its rationale to the Board, Meta determined that its original decision to take down the content was in error and it restored the post to the platform. The Board concludes that Meta's efforts to ensure respect for freedom of expression and assembly in the context of systematic state repression of freedom of expression have been insufficient and it recommends a change to the company's Crisis Policy Protocol.

Section 2: Case Description and Background

In July 2023, an Instagram user posted a video in Persian with English subtitles showing a man confronting a woman in public for not wearing the hijab, with the woman responding that she is standing up for her rights. The man is not identifiable in the video while the woman is fully visible. The video appears to be a repost of a recording initially shared by someone affiliated with or supporting the Iranian regime. The video was accompanied by a caption, also in Persian, expressing support for the woman and for Iranian women standing up to the regime, and criticizing the regime and its supporters. The caption included a phrase translated by Meta as, "it is not far to make you into pieces" and stated that the woman was arrested following the incident. The post had about 47,000 views, 2,000 likes, 100 comments and 50 shares.

This content was first flagged by an automated classifier, an algorithm Meta uses to identify potential violations of its policies, as potentially violating Instagram's Community Guidelines and sent for human review. Multiple reviewers assessed the content, but because of a technical error and because the reviewers did not reach the same conclusion on whether the post was a violation of the Violence and Incitement policy, it was initially not removed. A user then reported the content, in response to which an automated classifier determined again that the content potentially violated Meta's policies and sent it for additional review. The content was reported by one user and was reported only once. Following this additional level of review by Meta's team with regional and language expertise, Meta removed the post from Instagram under its Violence and Incitement policy. Meta's decision to remove the post was based on the following line in the caption: "it is not far to make you into pieces." The company read this line as a threat, targeting the man in the video who approached the woman for not wearing a hijab. The user who created the post appealed the decision to take down the post to Meta. A reviewer upheld the decision to remove.

The user who posted the content then appealed the removal to the Board. When the Board identified the case for legal review, Meta upheld its decision to remove the content. At this stage of review, Meta also considered the Coordinating Harm and Promoting Crime Community Standard, which prohibits "outing" unveiled women when this puts the woman at risk of harm. At the time the content was posted, the woman had already been arrested by the regime. After the Board selected the case, the company subsequently changed its decision and restored the content based on

additional input from its regional team and on the Board's decision in the Call for Women's Protest in Cuba case.

As the Board noted in its Iran Protest Slogan decision, people in Iran have been protesting against the government and for civil and political rights and gender equality, since at least the 1979 revolution. In 2023, the Nobel Peace Prize was awarded to Narges Mohammadi, an imprisoned human rights defender, for "more than 20 years of fighting for women's rights [which has] made her a symbol of freedom and standard-bearer in the struggle against the Iranian theocracy." Iran's criminal code penalizes women who appear in public without a "proper hijab" with imprisonment, a fine or lashes. Women in Iran are also banned from certain fields of study and many public places, and people are prohibited from dancing with members of the opposite sex, among other things. Men are considered the head of the household and women need the permission of their father or husband to work, marry or travel. A woman's court testimony is considered half the weight of a man's, which limits access to justice for women.

After Iranian authorities intensified and expanded mandatory hijab enforcement measures in 2022, women have faced increased scrutiny, often leading to verbal and physical harassment and arrests. In September 2022, 22-year-old Jina Mahsa Amini died in police custody three days after her arrest for allegedly failing to comply with the country's rules on wearing a "proper hijab." Her death sparked nationwide outrage and waves of protests across the country, and an anti-regime movement that has become known as: "Zan, Zendegi, Azadi" ("Woman, Life, Freedom"). This led to a violent crackdown by authorities, with over 500 confirmed deaths by the end of 2022, and an estimated 14,000 people being arrested, including protesters as well as journalists, lawyers, activists, artists and athletes who voiced support for the movement.

In September 2023, Iran's parliament approved a new "Hijab and Chastity" bill under which women could face up to 10 years in prison if they continue to defy the country's mandatory hijab rules. Businesses that serve women without a hijab would also face sanctions and risk being shut down.

Social media has been central to the women's protest movement in Iran, playing a critical role in the mobilization of protests and broadcasting of vital information (see

public comments PC 21007, PC-21011), and in documenting and publicly preserving evidence about abuses and human rights violations (PC-21008, attachment).

However, online campaigns also expose women to risks of further repression by the regime, including threats, defamation campaigns, arrests and imprisonment. Experts consulted by the Board noted an extensive network of entities associated with the Islamic Revolutionary Guard Corps and the Iranian government that operate on Instagram and Telegram, with the latter being frequently used to directly target and accuse protesters and dissenters.

Several public comments submitted to the Board also highlighted the regime's tactic of mass reporting protest content on Instagram using the user reporting system in order to "pressure social media companies into removing content related to dissidents or placing them into shadow bans," (see public comments PC-21011, PC-21009). There have also been reports of Iranian intelligence officials offering content moderators money to remove content shared by critics of the regime.

In February 2023, the UN Special Rapporteur on Iran reported concerns on the continuing repression and targeting of civil society activists, human rights defenders, women's rights activists, lawyers and journalists, as the authorities clamp down on avenues for expressing dissent, including heavy disruption of the internet and censorship of social media platforms.

Section 3: Oversight Board Authority and Scope

The Board has authority to review Meta's decision following an appeal from the person whose content was removed (Charter Article 2, Section 1; Bylaws Article 3, Section 1).

The Board may uphold or overturn Meta's decision (Charter Article 3, Section 5), and this decision is binding on the company (Charter Article 4). Meta must also assess the feasibility of applying its decision in respect of identical content with parallel context (Charter Article 4). The Board's decisions may include non-binding recommendations that Meta must respond to (Charter Article 3, Section 4; Article 4). When Meta commits to act on recommendations, the Board monitors their implementation.

When the Board selects cases like this one, in which Meta subsequently acknowledges that it made an error, the Board reviews the original decision to increase understanding of the content moderation process and to make recommendations to reduce errors and increase fairness for people who use Facebook and Instagram.

Section 4: Sources of Authority and Guidance

The following standards and precedents informed the Board's analysis in this case:

I. *Oversight Board Decisions*

The most relevant previous decisions of the Oversight Board include:

- Call for Women's Protest in Cuba
- Metaphorical Statement Against the President of Peru
- Iran Protest Slogan
- Öcalan's Isolation
- Breast Cancer Symptoms and Nudity

II. *Meta's Content Policies*

The Board's analysis was informed by Meta's commitment to voice, which the company describes as "paramount," and its values of safety, privacy and dignity.

Instagram Community Guidelines

The Instagram Community Guidelines state that the company will "remove content that contains credible threats" and links to the Violence and Incitement Community Standard. The Community Guidelines do not directly link to the Coordinating Harm and Promoting Crime Community Standard. Meta's Community Standards Enforcement Report for Q1 2023 states that "Facebook and Instagram share Content Policies. This means that if content is considered violating on Facebook, it is also considered violating on Instagram."

The content was removed under the Violence and Incitement policy. After the Board selected the case, Meta also analyzed the content under its Coordinating Harm and Promoting Crime policy.

Violence and Incitement Community Standard

According to the policy rationale, the Violence and Incitement Community Standard aims to "prevent potential offline violence that may be related to content" appearing on Meta's platforms. At the same time, Meta recognizes that "people commonly express disdain or disagreement by threatening or calling for violence in non-serious and casual ways." Meta therefore removes content when the company believes it contains "[t]hreats of violence that could lead to death," including content that targets anyone with "statements of intent" to commit "high-severity violence." It states that it considers the context of the statement when assessing whether a threat is credible, which can be any additional information such as the person's "public visibility and vulnerability of the target."

The "do not post" section of the policy specifically prohibits "threats of violence that could lead to death (and other forms of high-severity violence)." The word "threat" includes "statements of intent" to commit high-severity violence.

Coordinating Harm and Promoting Crime Community Standard

According to the policy rationale, the Coordinating Harm and Promoting Crime Community Standard aims to "disrupt offline harm and copycat behavior" by prohibiting people from "facilitating, organizing, promoting or admitting to certain criminal or harmful activities targeted at people, businesses, property or animals." This includes "outing," which Meta defines as content exposing the identity or locations affiliated with anyone who is alleged to, among other things, "be a member of an outing-risk group." This policy line is enforced by moderators at-scale in relation to certain specific groups. On escalation, with additional context, Meta's policy at the time the content was posted stated that it may also remove: "content that puts unveiled women at risk by revealing their images without [a] veil against their will or without permission." This language has now been edited to prohibit: "outing [unveiled women]: exposing the identity of a person and putting them at risk of harm."

III. Meta's Human Rights Responsibilities

The UN Guiding Principles on Business and Human Rights (UNGPs), endorsed by the UN Human Rights Council in 2011, establish a voluntary framework for the human rights responsibilities of private businesses. Meta's Corporate Human Rights Policy, announced on 16 March 2021, reaffirmed the company's commitment to respect rights as reflected in the UNGPs.

The following international standards may be relevant to the Board's analysis of Meta's human rights responsibilities in this case:

- The right to freedom of expression: Article 19, International Covenant on Civil and Political Rights (ICCPR), General Comment No. 34, Human Rights Committee, 2011; UN Special Rapporteur on freedom of opinion and expression, reports: A/HRC/38/35 (2018) and A/74/486 (2019); and Rabat Plan of Action, UN High Commissioner for Human Rights report: A/HRC/22/17/Add.4 (2013).
- The right to freedom of peaceful assembly: Article 21, ICCPR; General Comment No. 37, Human Rights Committee, 2020.
- The right to life: Article 6, ICCPR.
- The right to liberty and security of person: Article 9, ICCPR.
- The right to non-discrimination: Articles 2(1), 3 and 26, ICCPR; Article 1 and Article 7 (non-discrimination in participation in the political and public life of the country), Convention on the Elimination of All Forms of Discrimination against Women (CEDAW). See also Declaration on Human Rights Defenders, Article 8, right to effective access, on a non-discriminatory basis, to participation in the conduct of public affairs.
- The right to privacy: Article 17, ICCPR.

Section 5: User Submissions

The user who posted the content appealed the removal to the Board. In their statement, the user explained the post showed a representative of the Iranian government confronting a woman for not wearing a hijab. The user stated that the

video shows the bravery of the Iranian woman standing up for her rights. The user stated that others had shared similar videos on social media and that the content was not harmful or dangerous and did not violate any Instagram policy.

Section 6: Meta's Submissions

Violence and Incitement Community Standard

Meta told the Board that its initial decision to remove the content in this case was based on the Violence and Incitement policy. It explained that its decision was based on part of the post's caption, which the company translated as: "it is not far to make you into pieces." The company read that line as targeted towards the man in the video who approached the woman for not wearing a hijab. Its regional team determined that the phrase "make you into pieces" constituted a threat of physical harm in the Iranian context. Given that interpretation, Meta initially upheld the decision to remove the content under its Violence and Incitement policy.

Although Meta upheld its decision to remove the content when the Board first identified the case for legal review, it subsequently changed its decision after the Board selected this case and restored the content based on additional input from the regional team. Based on that input, Meta concluded that the most likely interpretation of the language in the caption was a reference to taking down the Iranian regime or those supporting the mandatory hijab, rather than a literal threat targeting the man in the video.

In this final review, Meta concluded that the content aimed to raise awareness and draw attention to abuses committed against women, such as the woman confronted by the man in the video for not wearing a hijab. The user refers to the strength of Iranian women and criticizes the "bastardness," referring to either the man filming the video or the regime as a whole, and raises awareness of the arrest of the woman in the video. Meta explained that the potentially threatening language should be read and understood in light of this overall context. This type of awareness raising, Meta noted in its rationale, citing the Board's decision in the Iran Protest Slogan case, is particularly important in Iran where there are limited outlets for free expression.

Meta also informed the Board that after additional research, regional teams suggested that the most reasonable interpretation of the language "it is not far to make you into pieces" was not a true threat, but rather a political critique directed toward either the regime as a whole or people who support the mandatory hijab requirement more generally. While "make you into pieces" generally refers to killing a person by cutting their body into pieces, here it could be understood as dismantling the regime (similar to the metaphorical language used in the Metaphorical Statement Against the President of Peru summary decision).

Meta told the Board that another factor in the company ultimately restoring the content was the Board's recent decision in the Call for Women's Protest in Cuba case, in which the Board emphasized that a contextual reading of the post should take into account the wave of state repression and the significant public interest in the historic protests that were the subject of the post. Additionally, the company considered the Iran Protest Slogan case, in which the Board analyzed the political movement around women's rights in Iran. There, the Board emphasized the importance of protecting voice in the context of the protest movement, particularly in light of the Iranian government's systematic repression of free expression and the importance of digital spaces as a forum to express dissent. The company also noted that it had considered the Board's recent summary decision in Metaphorical Statement Against the President of Peru, which re-emphasized the "importance of designing context-sensitive moderation systems with awareness to irony, satire or rhetorical discourse, especially to protect political speech."

Coordinating Harm and Promoting Crime Community Standard

Meta told the Board it also considered removing the content under the Coordinating Harm and Promoting Crime Community Standard for involuntarily outing the woman shown in the video without a veil. Meta enforces this policy line on escalation only and if additional context is provided. Enforcement requires input from relevant stakeholders and is focused on determining whether the content depicts an unveiled woman, exposing her identity without her permission, and is likely to put her at risk, rather than any specific terms used or the tone of the content or caption. Meta notes that a person cannot "out" themselves – outing must be involuntary to violate the policy.

In this case, Meta determined the content should not be removed for involuntary outing as the woman's identity was widely known and available online, and she had already been arrested at the time the case content was posted. This context significantly reduced the risk of harm associated with leaving the content on the platform.

The Board asked Meta 11 questions and two follow-up questions in writing. Questions related to enforcement procedures and resources for Iran, Meta's risk assessment for Iran in general and for the woman in the video in particular, automated and human review processes, the enforcement of content depicting unveiled women and the outing of an at-risk group, at-scale and on escalation. Meta answered all questions.

Section 7: Public Comments

The Oversight Board received 12 public comments relevant to this case. Seven of the comments were submitted from the United States and Canada, two from Central and South Asia, two from Europe and one from the Middle East and North Africa.

The submissions covered the following themes: the role of social media in the Iran protests, including the "Woman, Life, Freedom" movement, and the role that images of unveiled women play in digital campaigns; the risks for circulating imagery showing unveiled women in Iran on social media; the use of social media by the Iranian authorities; Meta's enforcement of its content moderation policies for Persian-language expression related to the political situation in Iran; freedom of expression, human rights, women's rights, government repression and social media bans in Iran.

To read public comments submitted for this case, please click here.

Section 8: Oversight Board Analysis

The Board examined Meta's original decision to remove the content under the company's content policies, human rights responsibilities and values.

The Board selected this case because it offered the opportunity to explore Meta's Violence and Incitement and Coordinating Harm and Promoting Crime policies, as well as related enforcement processes in the context of massive protests for women's

rights and women's participation in public life in Iran since September 2022. In particular, it addresses the importance of social media platforms for people protesting against the mandatory hijab rules.

Additionally, the case provides the Board with the opportunity to discuss Meta's internal procedures that determine when and why figurative speech, which, if understood literally, may be interpreted as threatening, but which given the context should not be interpreted as a credible threat. The case primarily falls into the Board's Elections and Civic Space priority, but also touches on the Board's priorities of Gender, Government Use of Meta's Platforms, and Crisis and Conflict Situations.

8.1 Compliance With Meta's Content Policies

I. Content Rules

Violence and Incitement Community Standard

The Board finds that the content in this case does not violate the Violence and Incitement Community Standard as it contains figurative speech expressing anger at government repression, rather than a literal and therefore credible threat of violence.

Meta explained that it originally removed the content in this case because it contained "a statement of intent to commit high-severity violence." It defines high-severity violence as a threat that could lead to death or is likely to be lethal.

Based on its regional team's assessment, Meta construed the phrase "it is not far to make you into pieces" from the post's caption as a threat of physical harm in the Iranian context, violating the Violence and Incitement policy.

Linguistic experts consulted by the Board explained that the relevant part of the caption can be translated as "we will tear you to pieces sometime soon!" or "it is not far away, we will rip you into shreds." The experts noted that the phrase in the Iranian context conveys anger, disappointment, resentment towards oppressors, and suggests the idea that the situation might eventually change as the oppressors' hold on power will not be eternal. This phrase should not be interpreted literally as an intention to cause physical harm; instead, it serves as a "rhetorical statement" aiming to attract attention,

emphasized by emotionally charged language featuring forceful verbs such as "tear" or "rip into pieces/shreds." These experts highlighted that such figurative speech mirrors the profound anger shared by both the user posting the content and their audience. Therefore, it does not imply actual threats of physical violence.

Although Meta told the Board that it also considers the context of the statement when assessing whether a threat is credible, its guidance to moderators does not indicate that they can consider context when assessing if there is a "statement of intent [to commit] high-severity violence." As long as the elements of the rule are satisfied, specifically when content includes both a threat and a target, the post is found violating, as was the case here. In this case, Meta interpreted the statement to be targeting the man following the woman. The rule does not require the target to be visible or identifiable. The rules provide only one example of threatening speech that is exempted or considered not to be a credible threat as a rule: "threats directed against certain violent actors, like terrorist groups."

The phrase "make you into pieces" here does not constitute a credible threat. Given the context of a period of turmoil, escalating repression and violence against people protesting the regime in Iran and considering the caption and the video as a whole, the Board finds that the phrase is figurative, not literal, expressing anger and dismay at the regime, and did not constitute a "statement of intent [to commit] high-severity violence." This interpretation is consistent with Meta's commitment to voice, and the importance of protecting expression of political discontent.

Coordinating Harm and Promoting Crime Community Standard

The Board finds that the content in this case does not violate the Coordinating Harm and Promoting Crime Community Standard.

Meta considered removing the post under the rule prohibiting "content that puts unveiled women at risk by revealing their images without [a] veil against their will or without permission." The policy line has since been edited to prohibit "Outing [unveiled women]: exposing the identity of a person and putting them at risk of harm." The company understands "outing" to include content that shares an image of an unveiled woman, exposes her identity without her permission and puts her at risk of harm. This

policy line is applied on escalation only and requires input from various stakeholders for enforcement (see Section 4 above).

In this case, the Board agrees with Meta that the content does not "out" the woman, as her identity was widely known and the risk of harm from the content had abated because she had already been arrested at the time the content was posted. Therefore, the video remaining on the platform would not meaningfully increase the level of risk to the woman and could in fact be protective in raising awareness of her case. Determining whether a post "outs" a woman and puts her at risk is especially context dependent; enforcing the policy on escalation-only can ensure the team enforcing it has the time and resources to effectively identify and consider the relevant context.

8.2 Compliance with Meta's Human Rights Responsibilities

The Board finds that removing the content from the platform was inconsistent with Meta's human rights responsibilities.

Freedom of Expression (Article 19 ICCPR)

Article 19 of the ICCPR provides for broad protection of expression, including "freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice." When restrictions on expression are imposed by a state, they must meet the requirements of legality, legitimate aim, and necessity and proportionality (Article 19, para. 3, ICCPR). The Board uses this framework to interpret Meta's voluntary human rights commitments, both in relation to the individual content decision under review and what this says about Meta's broader approach to content governance. As the UN Special Rapporteur on freedom of expression has stated, although "companies do not have the obligations of Governments, their impact is of a sort that requires them to assess the same kind of questions about protecting their users' right to freedom of expression," (A/74/486, para. 41).

Access to social media is crucial in a closed society such as Iran. As "digital gatekeepers," social media platforms have a "profound impact" on public access to information (A/HRC/50/29, para. 90). The post in this case is part of a broader protest movement that relies on digital civic spaces to survive. Laws determining how women

must dress impact their freedom and dignity (A/68/290, para. 38), whether the law seeks to prohibit wearing a veil or to proscribe going in public without one (see e.g., *Yaker v France,* CCPR/C/123/D/2747/2016). In this regard, "the Internet has become the new battleground in the struggle for women's rights, amplifying opportunities for women to express themselves," (A/76/258, para. 4). Empowering women's free expression enables their political participation and the realization of their human rights (A/HRC/Res/23/2, paras. 1-2; A/76/258, para. 5).

I. Legality (Clarity and Accessibility of the Rules)

The principle of legality requires any restriction on freedom of expression to be pursuant to an established rule, which is accessible and clear to users. The rule must be "formulated with sufficient precision to enable an individual to regulate his or her conduct accordingly and it must be accessible to the public," (General Comment No. 34, at para 25). Additionally, the rules restricting expression "may not confer unfettered discretion for the restriction of freedom of expression on those charged with [their] execution" and should "provide sufficient guidance to those charged with their execution to enable them to ascertain what sorts of expression are properly restricted and what sorts are not," (General Comment No. 34, at para 25; A/HRC/38/35 (undocs.org), at para 46). Lack of clarity or precision can lead to inconsistent and arbitrary enforcement of the rules. Applied to Meta, users should be able to predict the consequences of posting content on Facebook and Instagram and content reviewers should have clear guidance on their enforcement.

Violence and Incitement Community Standard

The Board finds that while the policy rationale – which expresses the aims of the Community Standard but is not part of the rule itself – of the Violence and Incitement Community Standard suggests that "context matters," and may be considered when evaluating a "credible threat," Meta's internal guidelines and approach to moderation do not enable this in practice. As the Board noted in the Iran Protest Slogan case, an at-scale content moderator is instructed to look for specific criteria, or elements, in the post, and once those elements are met, the moderators are instructed to remove the post. In other words, if the post contains a threat (e.g., "kill" or "I will tear you into pieces") and a target, the result is removal. Content moderators are not empowered to make an assessment on whether the threat is credible. As Meta also explained in that

case, the rote or formulaic approach is because "assessing whether [a phrase] constitutes rhetorical speech as opposed to a credible threat is challenging, particularly at scale." Consideration of credibility of threats is taken into account in creating the rule, not in enforcing it. As the Board noted in the Iran Protest Slogan case, while the "policy rationale appears to accommodate rhetorical speech of the kind that might be expected in protests contexts, the written rules and corresponding guidance to reviewers do not. Indeed, enforcement in practice, in particular at-scale, is more formulaic than the rules imply, and this may create misperceptions to users of how rules are likely to be enforced. The guidance to reviewers, as currently drafted, exclude[s] the possibility of contextual analysis, even when there are clear cues within the content itself that threatening language is rhetorical." This misalignment between the company's stated policy rationale and its actual enforcement practice continues and does not adequately satisfy the principle of legality.

The Board reiterates its findings from the Iran protest slogan case that Meta should provide nuanced guidance on how to take into account context, directing moderators to refrain from default removal of "rhetorical" or non-literal language expressing dissent, particularly in sensitive political environments, such as Iran.

Coordinating Harm and Promoting Crime Community Standard

The Board finds that Meta's prohibition on "content that puts unveiled women at risk by revealing their images without [a] veil against their will or without permission" is sufficiently clear, as applied in this case. It makes clear that content that "outs" unveiled women and could lead to harm is prohibited on Meta's platforms. However, the Board notes with concern that Instagram Community Guidelines do not directly link to the Coordinating Harm and Promoting Crime community standard. This undermines the accessibility of the rules to Instagram users. In previous cases (Breast Cancer Symptoms and Nudity, Öcalan's Isolation), the Board has recommended that Meta publicly clarify for users how Facebook Community Standards apply to Instagram. In response, Meta has undertaken a process to unify the Community Standards with Instagram's Community Guidelines and specify where policies differ slightly between platforms. In ensuing quarterly transparency reports, Meta has assured the Board that this effort remains a priority, noting that legal and regulatory

considerations have impacted their timelines. The Board reiterates the importance of moving quickly to complete this process and ensure clarity of applicable rules.

II. *Legitimate Aim*

Under Article 19, paragraph 3 of the ICCPR, expression may be restricted for a defined and limited list of reasons, including for the purpose of protecting the rights of others. In this case, the Board finds that the Violence and Incitement Community Standard aims to "prevent potential offline harm" by removing content that poses "a genuine risk of physical harm or direct threats to public safety." This policy therefore serves the legitimate aim of protecting the right to life (Article 6, ICCPR) and the right to physical security of a person (Article 9 ICCPR; General Comment No. 35, para. 9).

The Coordinating Harm and Promoting Crime policy serves the legitimate aim of protecting the rights of women in Iran to non-discrimination (Articles 2, 3 and 26, ICCPR; Articles 1 and 7, CEDAW), including in the enjoyment of their rights to freedom of expression and assembly (Articles 19 and 21, ICCPR), the right to take part in public life (Articles 1 and 7, CEDAW), the right to privacy (Article 17, ICCPR) and their rights to life (Articles 6, ICCPR) and to liberty and security of person (Article 9 ICCPR).

III. *Necessity and Proportionality*

Any restrictions on freedom of expression "must be appropriate to achieve their protective function; they must be the least intrusive instrument amongst those which might achieve their protective function; they must be proportionate to the interest to be protected" (General Comment 34, para. 34). Social media companies should consider a range of possible responses to problematic content beyond deletion to ensure restrictions are narrowly tailored (A/74/486, para. 51).

Violence and Incitement Community Standard

The Board finds that Meta's initial decision to remove the content in this case was not necessary. It was not required to protect the safety of the person taking the video or others, as the threat mentioned in the caption of the post is not literal. The Board is concerned that even after its guidance in the Iran Protest Slogan case, the company's Community Standards and guidance provided to moderators still leave room for

inconsistent enforcement of figurative (non-literal) threats, despite the situation in Iran, with protests ongoing for more than a year now. That case, like this one, involved a phrase that Meta failed to identify as a non-literal statement of a threat. Continuing lack of adequate guidance is further highlighted by the fact that the content in this case was reviewed by multiple at-scale moderators and teams within Meta, and repeatedly found to be violating.

As part of its analysis, the Board drew upon the six factors from the Rabat Plan of Action to evaluate the capacity of the content in this case to create a serious risk of inciting discrimination, violence or other lawless action. The Board notes that while the Rabat factors were developed for advocacy of national, racial or religious hatred that constitutes incitement, and not for incitement generally, the six-factor test is useful for assessing incitement in general terms, and the Board has used it in this way previously (see, for example, Iran Protest Slogan and Call for Women's Protest in Cuba):

- **Context:** The content was posted in July 2023, during a period of turmoil, escalating repressions and violence against people protesting the regime. There are limited venues for protest inside the country, given massive internet disruptions and the banning of major social media platforms like Facebook, Telegram and X. As Instagram is one of the few remaining platforms, its role in the "Woman, Life, Freedom" movement has been immeasurable. The experts consulted by the Board noted that the movement is aiming for normalization of protest and defiance against four decades of discriminatory laws against women. The experts observed that since the beginning of the movement, instilling fear and silencing women from expressing themselves online have been major trends. New laws and high-profile arrests are meant to curb these expressions of protest and stop the trend of normalizing defiance of hijab laws.
- **Identity of the speaker:** Meta told the Board that the company did not consider the user who posted the content to be a public figure. Based on the caption to the video, the user seems to be a supporter of or part of the "Woman, Life, Freedom" movement. Therefore, the speaker here is not in a position of authority and they are also potentially risking their own safety in voicing their support and posting this content.
- **Intent:** While assessing intent when moderating content at-scale presents a significant challenge, an objective and ordinary reading of the whole post indicates support for the

depicted woman and raising awareness about her arrest. According to the experts consulted by the Board, it is common in Iran for protesters to circulate the image of an unveiled woman and her name following her arrest, to put pressure on authorities to keep her safe. Protesters have learned that by publicizing these individuals, they can prevent further harassment of the victims.

- **Content and form of expression**: According to the linguistic experts consulted by the Board, the phrase "make you into pieces" in this context would be understood by Persian speakers as expressing a deep emotion, such as anger and disappointment, but not a literal threat of violence. Forceful and apparently threatening language has regularly been used by the movement to speak out against the regime. In the Iran Protest Slogan case, the Board found that the slogan "Death to Khamenei" constituted a "rhetorical threat" and noted that Meta had issued a "spirit of the policy" allowance for the phrase "I will kill whoever kills my sister/brother" for similar reasons. The phrase in this case appears in the middle of a caption praising a woman who is standing up for her rights. Considered in its entirety, the caption expresses support for Iranian women protesting discriminatory laws and abusive practices of the regime.

- **Extent and reach:** The post had about 47,000 views, 2,000 likes, 100 comments and 50 shares. Given that this speech does not appear to constitute incitement considering the other factors in the Rabat analysis, the high reach of the content in itself is not a factor indicating that removal was necessary.

- **Likelihood and imminence:** This factor assesses whether the speech is likely to trigger imminent and likely harm against the potential target of the speech, which in this case was the regime. The protest movement, and its supporters, are standing against a regime that is constantly using violent repression and reprisals against protesters. The most likely form of harm that can result from the speech in this case is retaliatory violence from the regime against the posting user or the woman shown in the video rather than violence against the regime or its supporters. Experts emphasized the dangers faced by individuals participating in these protests, and how the movement endures despite the looming threat of violence. They also stated that the aim of circulating the images of unveiled women who have been arrested is to call attention to their arrest and put pressure on the authorities to keep them safe.

Based on the analysis of the factors above, the Board considers that the content did not constitute a credible threat and was not capable of inciting offline harm. When figurative speech is used in the context of widespread protests met with violent repression, Meta should enable its reviewers to assess language and local context, aligning the guidance for moderators with the underlying policy rationale. Ensuring accurate assessment of whether a post is "figurative speech" or likely to incite violence is vital for improving moderation in crises more broadly. Automation accuracy will be impacted by the quality of the training data provided by human moderators. Where human moderators remove "figurative" statements due to a rigid enforcement of a rule, that mistake is likely to be reproduced and amplified through automation.

The Board notes that Meta has a number of mechanisms available to adjust its policies and their enforcement during crisis situations, including the "at risk" country tiering system, and its Crisis Policy Protocol. The "at risk" country tiering system is used to identify countries at risk of "offline harm and violence" in order to determine how the company should prioritize its product development or how to invest its resources. The assessment can also be taken into account for other processes (e.g., whether to stand up a special operations team or to trigger the use of its Crisis Policy Protocol). Meta informed the Board that for the second half of 2023, Iran has been designated an "at risk country." Iran has also been designated under the Crisis Policy Protocol since September 21, 2022, and has remained designated since that time. The Crisis Policy Protocol enables Meta to make certain temporary policy changes, known as "policy levers," to address a particular situation. Meta provided some examples of the policy levers already used in Iran, including "an allowance to permit content that includes the slogan 'I will kill whoever kills my sister/brother' or its derivatives, absent other violations of our policies." (For more examples of policy levers see Policy Forum Minutes, January 25, 2022, Crisis Policy Protocol.)

While the Board recognizes the company's commitment to safety and its efforts to mitigate potential content moderation risks by activating the Crisis Policy Protocol for Iran, these efforts have been insufficient to ensure respect for people's freedom of expression and assembly in an environment of systematic repression of dissent and social tensions. Research commissioned by the Board indicates the overwhelming majority of content depicting unveiled women in the context of discussing the wearing of hijab in Iran, on Meta's platforms, is shared by or in support of the protest movement.

Here, Meta's enforcement process repeatedly failed to distinguish figurative, or not literal, statements in relevant context from real threats and incitement to violence, which have the potential to further offline harm.

The Board recommends that Meta add a policy lever to the Crisis Policy Protocol, and accordingly provide internal criteria to at-scale moderators on how to identify statements using threatening language figuratively, or not literally, in the relevant context to be deemed non-violating of the Violence and Incitement policy line prohibiting threats of violence. In developing the crisis-specific criteria on how to determine whether a threat is figurative and not literal, Meta may look to the Rabat Plan factors (e.g., context of widespread protests against state repression, whether the speaker has the ability to incite or presents the risk of inciting people to engage in harm, relevant linguistic and social context indicating common use of strong/emotional language for rhetorical power, likelihood of harm given local knowledge, etc.). The company may also rely on its trusted partners to help devise or assess the criteria for moderation. Meta itself has stressed the practical importance of the Rabat Plan of Action to content moderation, having supported the United Nations in translating the Plan of Action into 32 languages. This policy lever should allow "figurative speech" within the context of protests against the regime, provided they are not intended to, and are not likely to, incite violence.

Coordinating Harm and Promoting Crime Community Standard

In this case, the Board finds removal under the Coordinating Harm and Promoting Crime Community Standard was not necessary, as the depicted woman's identity was widely known, and the content was clearly posted to call attention to her arrest, in the hope that the attention would lead to her release. Additionally, several public commentators highlighted that women who remove the hijab in public do it purposefully, as a form of protest, and are aware of the potential consequences, choosing "defiance as a strategic opposition to authority," (see public comment Tech Global Institute, PC-21009). The woman in the video had already been identified and arrested by the regime. This post was shared to call attention to that arrest. Protesters and dissidents detained by the regime have been tortured, subjected to gender-based violence or disappeared. Experts consulted by the Board and several public commentators specifically noted that this practice, calling attention to arrests and

calling for the release of a detained person, is regularly used by the movement and by human rights defenders in Iran, and can help protect individuals held by the regime.

Balancing the need to safeguard the identities of vulnerable users while avoiding censorship for those who desire exposure is a delicate question and requires contextual analysis, timely review and quick action.

Section 9: Oversight Board Decision

The Oversight Board overturns Meta's original decision to take down the content.

Section 10: Recommendations

A. Enforcement

To ensure respect for users' freedom of expression and assembly in an environment of systematic state repression, Meta should add a policy lever to the Crisis Policy Protocol providing that figurative (or not literal) statements, not intended to, and not likely to, incite violence, do not violate the Violence and Incitement policy line prohibiting threats of violence in relevant contexts. This should include developing criteria for at-scale moderators on how to identify such statements in the relevant context.

The Board will consider this recommendation implemented when Meta both shares with the Board the methods for implementing the policy lever and the resulting criteria for moderation in Iran.

***Procedural Note:**

The Oversight Board's decisions are prepared by panels of five Members and approved by a majority of the Board. Board decisions do not necessarily represent the personal views of all Members.

For this case decision, independent research was commissioned on behalf of the Board. The Board was assisted by an independent research institute headquartered at the University of Gothenburg which draws on a team of over 50 social scientists on six continents, as well as more than 3,200 country experts from around the world. The Board was also assisted by Duco

Advisors, an advisory firm focusing on the intersection of geopolitics, trust and safety, and technology. Memetica, an organization that engages in open-source research on social media trends, also provided analysis. Linguistic expertise was provided by Lionbridge Technologies, LLC, whose specialists are fluent in more than 350 languages and work from 5,000 cities across the world.