Call for women's protest in Cuba

2023-014-IG-UA

Case summary

The Oversight Board has overturned Meta's decision to remove a video posted by a Cuban news platform on Instagram in which a woman protests against the Cuban government, calls for other women to join her on the streets and criticizes men, by comparing them to animals culturally perceived as inferior, for failing to defend those who have been repressed. The Board finds the speech in the video to be a qualified behavioral statement that, under Meta's Hate Speech Community Standard, should be allowed. Furthermore, in countries where there are strong restrictions on people's rights to freedom of expression and peaceful assembly, it is critical that social media protects the users' voice, especially in times of political protest.

About the case

In July 2022, a news platform, which describes itself as critical of the Cuban government, posted a video on its verified Instagram account. The video shows a woman calling on other women to join her on the streets to protest against the government. At a certain point, she describes Cuban men as "rats" and "mares" carrying urinal pots, because they cannot be counted on to defend people being repressed by the government. A caption in Spanish accompanying the video includes hashtags that refer to the "dictatorship" and "regime" in Cuba, and it calls for international attention on the situation in the country, by using #SOSCuba.

The video was shared around the first anniversary of the nationwide protests that had taken place in July 2021 when Cubans took to the streets, in massive numbers, for their rights. State repression increased in response, continuing into 2022. The timing of the post was also significant because it was shared days after a young Cuban man was killed in an incident involving the police. The woman in the video appears to reference this when she mentions that "we cannot keep allowing the killing of our sons." Text overlaying the video connects political change to women's protests.

The video was played more than 90,000 times and shared fewer than 1,000 times.

Seven days after it was posted, a hostile speech classifier identified the content as potentially violating and sent it for human review. While a human moderator found the post violated

Meta's Hate Speech policy, the content remained online as it went through additional rounds of human review under the cross-check system. A seven-month gap between these rounds meant the post was removed in February 2023. On the same day in February, the user who shared the video appealed Meta's decision. Meta upheld its decision, without escalating the content to its policy or subject matter experts. A standard strike was applied to the Instagram account, but no feature limit.

Key findings

The Board finds that, when read as a whole, the post does not intend to dehumanize men based on their sex, trigger violence against them or exclude them from conversations about the Cuban protests. The post unambiguously aims to call attention to the woman's opinion about the behavior of Cuban men in the context of the historic demonstrations that began in July 2021. With the woman using language such as "rats" or "mares" to imply cowardice in that precise context, and to express her own personal frustration at their behavior, regional experts and public comments point to the post as a call-to-action to Cuban men.

If taken out of context and given an overly literal reading, the stated comparison of men to animals culturally perceived as inferior could be seen as violating Meta's Hate Speech policy. However, the post, when taken as a whole, is not a generalization that aims to dehumanize men, but instead a qualified behavioral statement, which is allowed under the policy. Consequently, the Boards finds that the removal of the content is inconsistent with Meta's Hate Speech policy.

Furthermore, with external experts flagging the hashtag #SOSCuba, posted by the user to draw attention to the economic, political and humanitarian crises facing Cubans, the protests are established as an important point of historical reference. The Board is concerned about how contextual information is factored into Meta's decisions on content that does benefit from additional human review. In this case, even though the content underwent escalated review–a process that is supposed to deliver better results–Meta still failed to get it right.

Meta should ensure that both its automated systems and content reviewers are able to factor contextual information into their decision-making process.

In this case, it was particularly important to protect the content. Cuba is characterized by closed civic spaces, so the risks associated with dissent are high, and access to internet is very restricted. In this case, relevant context may not have been sufficiently considered as part of the escalation process. Meta should consider how context influences its policies and the way in which they are enforced.

The Oversight Board's decision

The Oversight Board overturns Meta's decision to remove the post.

While the Board makes no new recommendations in this case, it reiterates relevant ones from previous decisions, for Meta to follow closely:

- Have a list-based over-enforcement prevention program to protect expression in line with Meta's human rights responsibilities, which should be distinct to the one that protects expression viewed by Meta as a business priority (recommendation no. 1 from the cross-check policy advisory opinion). This separate system should also ensure Meta provides additional layers of review to content posted by, among others, human rights defenders.
- Use specialized staff, with the benefit of local input, to create over-enforcement prevention lists (recommendation no. 8 from the cross-check policy advisory opinion).
- Improve how its workflow dedicated to meet its human rights responsibilities incorporates context and language expertise on enhanced review, specifically at decision-making levels (recommendation no. 3 from the cross-check policy advisory opinion).
- To ensure context is appropriately factored into content moderation, update guidance to its at-scale moderators with specific attention to rules around qualification, since the current guidance makes it virtually impossible for moderators to make the correct decisions (recommendation no. 2 from the Violence against women decision).

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

Full case decision

1. Decision summary

The Oversight Board overturns Meta's decision to remove an Instagram post published around the first anniversary of the historic nationwide protests that occurred in July 2021 in Cuba. In the post, a woman protests against the government and compares Cuban men to different animals that are culturally perceived as inferior. She does so to assert that Cuban men are not to be trusted because they have not acted with the forcefulness required to defend those who are being repressed. The post calls for women to hit the streets and

demonstrate to defend the lives of "our sons." Under Meta's Hate Speech policy, this is a *qualified behavioral statement* and, as such, should be allowed. In countries where there are strong restrictions on people's rights to freedom of expression and peaceful assembly, it is critical that social media protects the users' voice, especially in times of political protest.

2. Case description and background

In July 2022, a news platform's verified Instagram account, describing itself as critical of the government in Cuba, posted a video in which a woman calls on other women to join her in the streets to protest. A caption in Spanish includes quotes from the video, hashtags that refer to the "dictatorship" and "regime" in Cuba, and calls for international attention on the humanitarian situation in the country, including by using #SOSCuba. At one point in the video, the woman says that Cuban men are "rats" because they cannot be counted on to defend those who are being repressed by the government. At another point, she says that Cuban men are "mares" who carry urinal pots. The text overlaying the video connects political change to women's protests. The video was played more than 90,000 times and shared fewer than 1,000 times.

Public comments and experts familiar with the region, who the Board consulted, confirmed that these phrases are understood colloquially by Spanish speakers in Cuba to imply cowardice. One public comment (PC-13012) said that the terms, while insulting, "should not be interpreted as violent or dehumanizing speech." External experts said the term for "mares" is frequently employed as a homophobic insult or to refer to people as unintelligent. However, when combined with the reference to urinal pots, experts reported that the phrase "takes on the connotation that men are 'full of shit' and [is] utilized here to show women's discontent toward male figures" in the context of their inaction during political protests. In this sense, public comments point out that the woman does not disparage men by calling them "rats" or "mares," but that she uses this language to mobilize men in her country. According to these comments, men are not her enemies: she is just trying to awaken the conscience of men.

The post was shared around the first anniversary of the historic nationwide protests that occurred in July 2021 when Cubans took to the streets in what the Inter-American Commission on Human Rights (IACHR) described as "a peaceful protest to claim their civil

liberties and demand changes to the country's political structure." The IACHR reported that Cubans "were also protesting the lack of access to economic, social, and cultural rights – especially because of persistent food and medicine shortages and the escalating impacts of the COVID-19 pandemic. According to civil society and international bodies – such as the European Parliament – the massive protest of July 11 was among the largest demonstrations in Cuba's recent history. These protests triggered immediate state reactions against the demonstrators" (Inter-American Commission of Human Rights, 2022 Annual Report, para. 43). From July 2021 onwards and throughout 2022, state repression increased. The post was published in the context of this significant social tension. Additionally, it was shared days after a young Cuban man was killed in an incident involving the police. Some parts of that incident were documented on social media, and the woman speaking in the video appears to reference this when she says: "we cannot keep allowing the killing of our sons." External experts who analyzed the social-media response found a broader pattern of users referring to the teenager's killing as a way to articulate their criticism of the government and to call for civilian action: "the discourse in the comment sections of the largest Instagram posts centered around the common themes of dictatorship, police brutality, and the lack of action from bystanders."

External experts familiar with the region highlighted the importance of social-media campaigns that use hashtags such as #SOSCuba in raising awareness around the economic, political, and humanitarian crises faced by Cubans. In the wake of the 2021 protests, the government intensified its crackdown on virtually all forms of dissent and public criticism. The IACHR documented eight waves of repression by the Cuban state in which it observed "(1) the use of force and intimidation and smear campaigns; (2) arbitrary arrests, mistreatment, and deplorable prison conditions; (3) criminalization of protesters, judicial persecution, and violations of due process; (4) closure of democratic forums through repression and intimidation to discourage new social demonstrations; (5) ongoing incarceration, trials without due process guarantees, and harsh sentences; (6) legislative proposals aimed at curtailing, surveilling, and punishing dissent and criticism of the Government and at criminalizing the actions of independent civil society organizations; (7) harassment of relatives of persons detained and charged for taking part in the protests; and (8) deliberate cuts in Internet access" (IACHR, 2022 Annual Report, para. 44). The IACHR noted that, although the waves of repression began in the second half of 2021, they continued throughout 2022, and that dozens of people were injured by police through the disproportionate use of force (IACHR, 2022 Annual Report, para. 46). On July 11, 2022, the

IACHR and its special rapporteurs condemned the persistent state repression of 2022 that occurred in response to the demonstrations of 2021.

The legislative response to the July 2021 protests also included further criminalization of online speech, including new penal code regulation establishing heightened penalties for alleged offenses such as spreading "fake information" or offending someone's "honor" on social media, or in online or offline media. This is supplementary to existing provisions of the penal code, which cover "public disorder," "resistance," and "contempt," and have historically been used to stifle dissent and criminalize protests. According to the IACHR, "the new text imposes harsher penalties and uses broad, imprecise language to define offenses, such as sedition and crimes against constitutional order" (IACHR, 2022 Annual Report, para. 97). Despite these displays of force and legal actions by the government after July 2021, external experts familiar with the region documented several attempts to organize localized protests against the government, but noted the significant risks of participation.

Near-complete government control of the internet's technical infrastructure in Cuba, in addition to censorship, obstruction of communications, and the very high cost of accessing the internet, "prevents all but a small fraction of Cubans from reading independent news website and blogs" (IACHR, 2022 Annual Report, para. 69). The Board also makes note of the attempts by government-linked networks described by Meta in its February 2023 report on Adversarial Threat to "create the perception of widespread support for the Cuban government across many internet platforms, including Facebook, Instagram, Telegram, Twitter, YouTube and Picta, a Cuban social network." According to Meta, the company's investigation found links between the Cuban government and the people behind a network of 363 Facebook accounts, 270 pages, 229 groups and 72 accounts on Instagram, which violated Meta's policy against coordinated inauthentic behavior.

Seven days after the video was posted on the Instagram account in July 2022, a hostile speech classifier identified the content as potentially violating and sent it for human review. The following day, a human moderator reviewed the content and found the post violated Meta's Hate Speech policy. Meta did not consider the woman depicted in the video to be a public figure. Based on the account's cross-check status, the content in this case was then escalated for secondary review. The first moderator in the secondary review process assessed the content as violating on July 12, 2022. The second moderator assessed the content as violating on February 24, 2023. Meta then removed the content from Instagram on the same

day, more than seven months after it was initially flagged by the company's automated systems. The delay in the review was caused by a backlog in Meta's review queues under the cross-check system.

On the same day the content was removed, the user who shared the video appealed Meta's decision. The content was again reviewed by a moderator who, on February 26, 2023, upheld the original decision to remove it. The content was not escalated to policy or subject matter experts for additional review at this time. According to Meta, a standard strike was applied to the user's account. However, no feature limit was applied to the account in line with Meta's account restriction protocols. The user then appealed the case to the Board.

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 the Board's decision in respect to 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.

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:

- Violence against women cases (2023-002-IG-UA; 2023-005-IG-UA)
- Iran protest slogan case (2022-013-FB-UA)
- Knin cartoon case (2022-001-FB-UA)
- South Africa slurs case (2021-011-FB-UA)

- Colombia protests case (2021-010-FB-UA)
- Pro-Navalny protests in Russia case (2021-004-FB-UA)
- Depiction of Zwarte Piet case (2021-002-FB-UA)
- Meta's cross-check program (PAO-2021-02)

II. *Meta's content policies:*

The Instagram Community Guidelines state that content containing hate speech will be removed. Under the heading "Respect other members of the Instagram community," the guidelines state that it is "never OK to encourage violence or attack anyone based on their race, ethnicity, national origin, sex, gender, gender identity, sexual orientation, religious affiliation, disabilities, or diseases." The Instagram Community Guidelines then link the words "hate speech" to the Facebook Hate Speech Community Standard.

The Hate Speech policy rationale defines hate speech as a direct attack against people on the basis of protected characteristics, including sex, gender, and national origin. Meta does not allow Hate Speech on its platform because it "creates an environment of intimidation and exclusion, and in some cases may promote offline violence." The rules prohibit "violent" or "dehumanizing" speech against people based on these characteristics, including men.

Tier 1 of the Hate Speech policy prohibits "dehumanizing speech or imagery in the form of comparisons, generalizations, or unqualified behavioral statements (in written or visual form) to or about [...] [a]nimals in general or specific types of animals that are culturally perceived as intellectually or physically inferior." Additionally, Meta's internal guidelines to content reviewers on how to apply the policy define "qualified" and "unqualified" behavioral statements and provide examples. Under these guidelines, "qualified statements" do not violate the policy, while "unqualified statements" are violating and removed. Meta states qualified behavioral statements use statistics, reference individuals, or describe direct experience. Meta also states that, under the Hate Speech policy, the company allows people to post content containing qualified behavioral statements about protected characteristic groups when the statement discusses a specific historical event (for example, by referencing statistics or patterns). According to Meta, unqualified behavioral statements "explicitly attribute a behavior to all or a majority of people defined by a protected characteristic."

The Board's analysis was informed by Meta's commitment to "Voice," which the company describes as "paramount," and its values of "Safety" and "Dignity."

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. In 2021, Meta announced its Corporate Human Rights Policy, in which it reaffirmed its commitment to respecting human rights in accordance with the UNGPs.

The Board's analysis of Meta's human rights responsibilities in this case was informed by the following international standards:

- The rights to freedom of opinion and expression: Articles 19 and 20, International Covenant on Civil and Political Rights (ICCPR); General Comment No. 34, Human Rights Committee, 2011; UN Special Rapporteur (UNSR) on freedom of opinion and expression, reports: A/HRC/38/35 (2018), A/74/486(2019), A/76/258 (2021); 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 non-discrimination: Article 2, para. 1 and Article 26, ICCPR.

5. User submissions

In their appeal to the Board, the content creator called on social-media companies to better understand the "critical situation" in Cuba, flagging that the video makes references to the July 2021 protests. The content creator also explained that the woman in the video is calling on Cuban men to "do something to solve" the crisis.

6. Meta's submissions

Meta removed the post under Tier 1 of its Hate Speech Community Standard because it attacked men by comparing them to rats and horses carrying human waste. The company explained that rats are a "classic example" of "animals that are culturally perceived as intellectually or physically inferior." While the company is not aware of any specific trope or

cultural tradition associated with “mares loaded with chamber pots or toilets,” the phrase violates the Hate Speech policy, according to Meta, because it compares men to the “repulsive image of animals that are presumably carrying human urine and feces.”

Meta explained that “the comparisons to rats and toilet-laden horses dehumanizes men based on their sex.” Meta also said that “this excludes men from the conversation and could result in them feeling silenced.”

In its response to the Board’s questions, Meta stated that the company considered applying a “spirit of the policy” allowance to this content. Meta makes “spirit of the policy” exceptions to allow content when a strict application of the relevant Community Standard produces results that are inconsistent with its rationale and objectives. However, Meta concluded that such an allowance was not appropriate because the content violates both the letter and the spirit of the policy.

Meta further explained that under the Hate Speech Community Standard, it treats all groups defined by protected characteristics equally. According to the company, violating hate speech attacks by one marginalized protected characteristic group directed at another protected characteristic group will be removed. Meta explained that as part of its Hate Speech policy, the company approaches all protected characteristic groups in the same way, so that globally they receive equitable treatment and so the policy can be enforced at scale. Meta refers to this approach as being “protected characteristic-agnostic.” Meta stated that when content is escalated for additional review by human moderators, it does not allow hate speech or “spirit of the policy” allowances based on asymmetrical power dynamics (i.e., when the target of the hate speech is a more powerful group) “for the same reason we have a protected characteristic-agnostic policy.” Meta stated that it “cannot and should not rank which protected characteristic groups are more marginalized than others.” Instead, Meta focuses on “whether there is an attack against a group of people based on their protected characteristics.” Meta acknowledged that some stakeholders have said the Hate Speech policy should differentiate between content that is perceived to be “punching down,” which should be removed, versus content that is “punching up,” which should be allowed because it may imply themes of social justice. However, Meta said that “there is little consensus among stakeholders about what counts as ‘punching down vs. punching up.’”

The Board also asked how contextual information, asymmetrical power dynamics between protected characteristic groups, and information about the political environment in which a post is made factor into the hostile speech classifier's decision to send content for human review. In response, Meta said that "the context that a classifier takes into account is within the post itself" and that it "does not consider other contextual information from global events." In this case, the hostile speech classifier identified the content as potentially violating Meta's policies and sent it for human review.

The Board asked 17 questions in writing. The questions addressed issues relating to Meta's content-moderation approach in Cuba; the bearing that asymmetrical power dynamics have on the Hate Speech Community Standard, as well as its enforcement following automated and human review; and opportunities for context assessment, specifically within the part of Meta's cross-check system called Early Response Secondary Review (ERSR). ERSR is a type of cross-check that provides additional levels of human review for certain posts initially identified as violating Meta's policies while keeping the content online. All 17 questions were answered by Meta.

7. Public comments

The Oversight Board received 19 public comments relevant to this case. Nine of the comments were submitted from United States and Canada; three from Latin America and Caribbean; five from Europe; one from Asia Pacific; and one from the Middle East and North Africa. The submissions covered the following themes: the human rights situation in Cuba; the importance of an approach to content moderation that recognizes linguistic, cultural, and political nuances in calls for protest; gender-based power asymmetries in Cuba; the intersection of hate speech and calls for protest; and online and offline protest dynamics in Cuba.

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

8. Oversight Board analysis

The Board examined whether this content should be restored by analyzing Meta's content policies, human rights responsibilities and values. The Board also assessed the implications of this case for Meta's broader approach to content governance.

The Board selected this appeal because it provides an opportunity to better understand how Meta's Hate Speech policy and its enforcement impact calls for protest in contexts characterized by restricted civic spaces.

8.1 Compliance with Meta's content policies

I. *Content rules*

The Board finds that the content in this case is not hate speech as per Meta's Community Standards, but a *qualified behavioral statement* and, as such, is allowed under the Hate Speech policy. Consequently, the removal of the content is inconsistent with this policy. It is true that the statements in which men are compared with "rats" or "mares" loaded with urinal pots, in a literal reading and out of context, could be interpreted as violating Meta's policy on hate speech. Nevertheless, the post, taken as a whole, is not a generalization that aims at dehumanizing or triggering violence against all men, or even the majority of men. The Board finds that the statements directed at men are *qualified* in the sense they unambiguously aim at calling attention to the behavior of Cuban men in the context of the historic demonstrations that began in July 2021 in Cuba, and which were followed by state repression that continued into 2022 in response to subsequent calls for protest. The content creator explicitly refers to those events in the post by using the #SOSCuba hashtag. The content is a commentary on how an identifiable group of people have acted, not a statement about character flaws that are inherent in a group.

According to public comments and experts consulted by the Board, epithets such as "rats" or "mares" are used in the vernacular Spanish spoken in Cuba in heated discussions to imply cowardice. As such, the terms should not be read literally and do not indicate that men have inherently negative characteristics by virtue of being men. Rather, they mean that Cuban men

have not acted with the necessary forcefulness to defend those who are being repressed by the government in the context of the protests.

The post was shared in the context of a wave of state repression that took place around the first anniversary of the historic nationwide protests, which occurred in July 2021. External experts have flagged the hashtag #SOSCuba, used by the content creator, as an important one adopted by social-network campaigns to draw attention to the economic, political, and humanitarian crises facing Cubans. The use of the hashtag in addition to the woman's warning statement in the video ("we cannot keep allowing the killing of our sons") and her call "to the streets," demonstrate how the events that began in July 2021 are established as an important point of historical reference for subsequent efforts by citizens to mobilize around social and political issues that continued into 2022. Therefore, the post reflects the user's opinion about the behavior of a defined group of people, Cuban men, in the specific context of a historical event.

In conclusion, the Board finds that, when read as a whole, the post does not intend to dehumanize men, generate violence against them or exclude them from conversations about the Cuban protests. On the contrary, the woman in the video is questioning what, in her opinion, the behavior of Cuban men has been in the precise context of the protests, and she aims to galvanize them to participate in such historic events. The content in this case is, therefore, a statement of qualified behavior on an issue of significant public interest related to the historic protests and the wave of repression that followed.

In response to this case, a minority of the Board questioned the agnostic enforcement of Meta's Hate Speech policy, particularly in situations when such enforcement can lead to further silencing of historically marginalized groups. For these Board Members, a proportionate Hate Speech policy should acknowledge the existence of power asymmetries when such acknowledgment can prevent the suppression of under-represented voices.

Finally, the Board agrees that the post falls directly within Meta's paramount value of "Voice." Therefore, its removal was not consistent with Meta's values. A similar approach was taken by the Board in relation to one of the posts reviewed in the Violence against women cases, when the Board agreed with Meta's ultimate conclusion that the content should be taken as a whole and assessed as a qualified behavioral statement.

II. *Enforcement action*

According to Meta, after a hostile speech classifier identified the content as potentially violating Meta's policies, it was sent for human review. Between the first human review and first level of secondary review on July 12, 2022, both of whom found the content to violate Meta's Hate Speech policies, and the second level of secondary review on February 24, 2023, when an additional moderator found the content violating and removed the post, more than seven months elapsed. As described in Section 2, the delay in the review was caused by a backlog in Meta's cross-check system. As part of the Board's cross-check policy advisory opinion, Meta disclosed that the cross-check system had been operating with a backlog of content that delays decisions. In information that Meta provided to the Board, the longest time a piece of content remained in the ERSR queue was 222 days; the delay of more than seven months observed in this case is similar to this length. According to Meta, as of June 13, 2023, the review of backlogged content in the ERSR program queue has been completed in response to recommendation no. 18 from the cross-check policy advisory opinion, which said that Meta should not operate this program with a backlog.

The Board notes the seven-month delay in this case. The delay ultimately meant the content remained on the platform while waiting for the final stage of cross-check secondary review. The content remaining on the platform is an outcome in line with the Board's analysis of the application of the Hate Speech Community Standard. However that outcome was not in accordance with Meta's understanding that the content was harmful.

The enforcement history in this case also raises concerns about how contextual information is factored into decisions on content that does benefit from additional human review. The Board has previously acknowledged that assessing the use of hate speech and relevant context at scale is a difficult challenge (see Knin cartoon case). In particular, the Board has emphasized that dehumanizing discourse, through implicit or explicit discriminatory acts or speech, has, in some circumstances, resulted in atrocities (see Knin cartoon case). The Board has also considered that, in certain circumstances, moderating content with the objective of addressing cumulative harms caused by hate speech at scale may be consistent with Meta's human rights responsibilities, even when specific pieces of content, seen in isolation, do not appear to directly incite violence or discrimination (see Depiction of Zwarte Piet case).

In order to avoid inappropriately stifling public debate on highly relevant issues, such as violence against women (see Violence against women cases) or, as in this case, political speech on historical events, Meta has established exceptions such as the one on qualified behavioral statements. Making sure content reviewers are able to accurately distinguish between qualified and unqualified behavior statements is therefore necessary for Meta to reduce false positive (mistaken removal of content that does not violate its policies) rates in the enforcement of the Hate Speech policy. For the same reason, it is important for Meta to ensure that both its automated systems, including content machine learning classifiers that screen for what Meta considers "hostile speech," and human content reviewers are able to factor contextual information into their determinations and decisions. This is especially important to reiterate when, as in this case, Meta's content reviewers do not take context into account and remove a post when it is particularly urgent to protect it. Indeed, operational mechanisms and processes aimed at surfacing contextual insights are especially significant for countries or regions characterized by closed civic spaces, where the risks associated with dissent and criticism of the government are much higher, and access to internet is very restricted. The Board also notes that reviews at escalation level are supposed to deliver better results, even in difficult cases, since better tools for assessing context are available. However, even after the content in this case underwent escalated review, Meta still failed to get it right and keep the post on Instagram.

As part of the cross-check policy advisory opinion, Meta explained that, generally for ERSR, the markets team (which includes a mix of Meta full-time employees and full-time contractors) first reviews the content. This team has additional contextual and language knowledge about a specific geographic market. According to Meta, the Cuban market "is not a separate market and it is categorized in [Meta's] general Spanish language ESLA queues (Español Latin)," meaning that content from Cuba is reviewed by reviewers covering Spanish-language content in general and not focusing specifically on that country. Meta said that "other countries are split" in queues for country-specific or region-specific review (e.g. Spain queues for Spain, VeCAM (Venezuela, Honduras, Nicaragua) queues for Venezuela and Central America)." The Early Response team (an escalations team comprising Meta full-time employees only) may then review to confirm whether the content is violating. According to Meta, this team has "deeper policy expertise and the ability to factor in additional context" and may also apply Meta's "newsworthiness" and "spirit of the policy" allowances. However, to assess the content, the Early Response team relies on translations and contextual

information provided by the relevant Regional Market team and does not have language or regional expertise.

Given Meta's decision in this case, the Board is concerned that relevant contextual information – such as the whole content of the post, the #SOSCuba hashtag, the events around the one-year anniversary of the historic July 2021 protests, the wave of repression denounced by international organizations at the time the post was published and, among other things, the death of a young Cuban in an incident involving the police – may not have been sufficiently considered when assessing the content as part of the cross-check escalations process.

In response, the Board reiterates recommendation no. 3 from the cross-check policy advisory opinion, which called on Meta to "improve how its workflow dedicated to meet Meta's human rights responsibilities incorporates context and language expertise on enhanced review, specifically at decision making levels." Meta has agreed to fully implement this recommendation. In Meta's Q1 2023 update, the company stated it has already taken certain initiatives to incorporate context and language expertise at the ERSR level. The Board hopes that context and language expertise would help prevent future content like the post considered here from being removed.

8.2 Compliance with Meta's human rights responsibilities

The Board finds that Meta's decision to remove the content in this case 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 about politics, public affairs, and human rights, with expression about social or political concerns receiving heightened protection (General Comment No. 34, paras. 11-12). Article 21 of the ICCPR provides protection for freedom of peaceful assembly – and assemblies with a political message are accorded heightened protection (General Comment No. 37, paras. 32 and 49). Extreme restrictions on freedom of expression and assembly in Cuba make it especially crucial that Meta respect these rights, particularly in times of protest (Colombia protests

decision; Iran protest slogan decision; General Comment No. 37, para. 31). Article 21's protection extends to associated activities that take place online (Ibid., paras. 6 and 34). As highlighted by the UN Special Rapporteur (UNSR) on the right to freedom of expression, "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).

The expression at issue in this case deserves "heightened protection" because it involves a woman's call for protest to defend the rights of those who have been repressed, one which came at a significant political moment, almost one year after historic protests in Cuba in July 2021. Public anger and criticism of the Cuban government continued as Cuban authorities intensified their legal and physical crackdowns on expressions of dissent in the year following the July 2021 protests. According to experts, while those sentiments can manifest as smaller protests in response to local events (such as the death of the Cuban teenager in this case), the persistence of citizens' concerns around the economy, governance, and fundamental freedoms, combined with internet connectivity (albeit constrained by high costs and state control of important infrastructure), have made it clear that protests are "here to stay."

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). These requirements are often referred to as the "three-part test." 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 UNSR 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).

I. Legality (clarity and accessibility of the rules)

The principle of legality requires rules that limit expression to be clear and publicly accessible (General Comment No. 34, para. 25). The Human Rights Committee has further noted that rules "may not confer unfettered discretion for the restriction of freedom of expression on those charged with [their] execution" (*Ibid.*). In the context of online speech, the UNSR on freedom of expression has stated that rules should be specific and clear (A/HRC/38/35, para.

46). People using Meta's platforms should be able to access and understand the rules, and content reviewers should have clear guidance on their enforcement.

Meta's Hate Speech policy prohibits content attacking groups on the basis of protected characteristics. Meta defines attacks as "violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation." Dehumanizing speech includes comparisons, generalizations, or unqualified behavioral statements about or to animals culturally perceived as inferior. The same policy, however, allows *qualified behavioral statements*. Meta's enforcement error in this case demonstrates that the policy's language and the internal guidance provided to content reviewers are not sufficiently clear in order for reviewers to accurately determine when a qualified behavioral statement has been made.

According to Meta, unqualified behavioral statements "explicitly attribute a behavior to all or a majority of people defined by a protected characteristic." Meta further explained that the company allows qualified behavioral statements about protected characteristic groups when the statement discusses a specific historical event (for example, by referencing statistics or patterns). In the Violence against women case, Meta informed the Board that "it can be difficult for at-scale content reviewers to distinguish between qualified and unqualified behavioral statements without taking a careful reading of context into account." However, the guidance to reviewers, as currently drafted, significantly limits their ability to perform an adequate contextual analysis, even when there are clear cues within the content itself that it includes a qualified behavioral statement. Indeed, Meta stated that because it is challenging to determine intent at scale, its internal guidelines instruct reviewers to default to removing behavioral statements about protected characteristic groups when the user has not made it clear whether the statement is qualified or unqualified.

In the present case, the post, read as a whole, unambiguously reflects the critical judgment of the woman in the video when she refers to the behavior of Cuban men in the specific context of the historic Cuban protests of 2021 and the wave of repression that followed in 2022. The whole content, including the hashtag #SOSCuba, and the events publicly known at the time of publication, make it clear that the post was, in fact, a statement discussing specific historical and conflict events through the reference to what the woman in the video understands as a pattern.

As discussed in the Violence against women decision and in the Knin cartoon decision, content reviewers should have sufficient opportunities and resources to take contextual cues into account in order to accurately apply Meta's policies. The Board finds that the language of the policy itself and the internal guidelines to content reviewers are not sufficiently clear to ensure that qualified behavioral statements are not wrongfully removed. The company's confusing, or even contradictory, guidance makes it difficult for reviewers to reach a reliable, consistent and predictable conclusion. The Board reiterates recommendation no. 2 from the Violence against women decision, which urged Meta to "update guidance to its at-scale moderators with specific attention to rules around qualification."

II. Legitimate aim

Any restriction on expression should pursue one of the legitimate aims listed in the ICCPR, which include the "rights of others." In several decisions, the Board has found that Meta's Hate Speech policy, which aims to protect people from the harm caused by hate speech, has a legitimate aim that is recognized by international human rights law standards (see, for example, Knin cartoon decision).

III. Necessity and proportionality

The principle of necessity and proportionality provides that 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; [and] they must be proportionate to the interest to be protected" (General Comment No. 34, para. 34). While the Board finds the content in this case is not hate speech and should remain on Instagram, the Board is not indifferent to the difficulties of moderating hate speech that includes comparisons to animals (see Knin cartoon decision).The UNSR on freedom of expression has noted that on social media, "the scale and complexity of addressing hateful expression presents long-term challenges" (A/HRC/38/35, para. 28). The Board, relying on the Special Rapporteur's guidance, has previously explained that, although these restrictions would generally not be consistent with governmental human rights obligations (particularly if enforced through criminal or civil penalties), Meta may moderate such speech if it demonstrates the necessity and proportionality of the speech restriction (see South Africa slurs decision). In the event of inconsistencies between company rules and international standards, the Special Rapporteur has called on social-media companies to "give a reasoned

explanation of the policy difference in advance, in a way that articulates the variation" (A/74/486, para. 48).

As previously mentioned in Section 8.1, Meta's hate speech policy contains several exceptions, one of which is precisely at issue in this case: qualified statements about behavior.

Meta understood that no exception was applicable and removed the content. The Board, however, found that in applying an overly literal reading of the content, Meta overlooked important context; disregarded a relevant carve-out from its own policy; and adopted a decision that was neither necessary nor proportionate to achieve the legitimate aim of the Hate Speech policy.

In this case, the Board considered the Rabat Plan factors in its analysis (OHCHR, A/HRC/22/17/Add.4, 2013) and took into account the differences between the international law obligations of states and the human rights responsibilities of Meta, as a social media company. In its analysis, the Board focused on the social and political context, the author, the content itself and form of the speech.

As previously mentioned in this decision, the post was published in the context of high social tension characterized by a strong wave of repression arising from the historic protests in Cuba that began in 2021. The Board also notes the death of a young Cuban man in an incident involving the police as relevant context, as it catalyzed calls for protest against the government, such as the one in this case's content. In the post, a woman issues a statement about what, in her opinion, has been the behavior of Cuban men during the protests and calls on women to take to the streets to defend the lives of "our sons." The post includes explicit references to the protests and the #SOSCuba hashtag. Linguistic analysis of the post in its entirety and in the context in which it was published leaves no doubt as to its meaning and scope. The post does not attribute a behavior to all men nor to the majority of men. Nor does it purport to or contribute to dehumanizing all or most of a protected characteristic group. The post does not generate violence towards men, nor does it exclude them from public conversations. On the contrary, amid high social tension, it resorts to strong language to encourage Cuban men to participate in protests by saying they have not lived up to their responsibilities. However, despite the fact that the content does not contribute to the

generation of any harm, its removal has a significant negative impact on the woman depicted in the video, on the user who shared it and, ultimately, on the political debate.

Indeed, Meta's decision to remove the post is likely to have had a disproportionate impact on the woman in the video who overcame many difficulties that exist in Cuba, including access to the internet and the risks of speaking out against the government. Additionally, the removal is likely to have placed an unnecessary burden on the user – the news platform – which has had to overcome barriers to disseminate information about what is happening in Cuba. The strike Meta applied to the user's account following the post's removal could have aggravated the situation, and potentially resulted in the account's suspension. Finally, the Board also finds that the post is in the public interest and contains a call for protest that is passionate, but does not advocate violence. Therefore, the post's removal also impacts the public debate in a place where it is already severely limited.

The UNSR on freedom of expression has stated in relation to hate speech that the "evaluation of context may lead to a decision to make an exception in some instances, when the content must be protected as, for example, political speech" (A/74/486, para. 47 (d)).

The Board has repeatedly affirmed the importance of this assertion. In the Colombia protests decision, the Board examined the challenges of assessing the political relevance and public interest of content containing a homophobic slur within a protest context. The Iran protest slogan decision acknowledged that "Meta's current position is leading to over-removal of political expression in Iran at a historic moment and potentially creates more risks to human rights than it mitigates." Finally, beyond contextual signals within the content itself, in the Pro-Navalny protests in Russia decision, the Board affirmed the importance of external context, saying, "context is key for assessing necessity and proportionality . . . Facebook should have considered the environment for freedom of expression in Russia generally and specifically government campaigns of disinformation against opponents and their supporters, including in the context of the January protests." While that case concerned Meta's Bullying and Harassment policy, the observations on the "environment for freedom of expression" and protests apply to this case on hate speech, too.

The Board notes the significant constraints on freedom of expression in Cuba, as well as the physical and legal risks that come with speaking against the government (Section 2). These risks, along with the high cost of data and internet access in Cuba, raise the stakes of

moderating content from dissenting voices in the country. One public comment (PC-13017) highlighted the importance of "safeguard[ing] the limited avenues for dissent and organization of protests."

Finally, the Board considered the IACHR's 2022 report, which notes that the Commission was "informed of persecution, political violence, and sexual assaults against women by state agents in the context of social protests; this is reported to be even more severe in the case of female human rights activists and defenders" (IACHR, 2022 Annual Report, para. 166). Independent media coverage about Cuba has also highlighted the impact of government responses to the July 2021 protests on women, with some civil society organizations arguing that "the greatest manifestation of gender violence in the Cuban context is by the government, and is explicitly demonstrated with the update of the list of women deprived of their freedom for political reasons."

The Board urges Meta to exercise more care when assessing content from geographic contexts where political expression and peaceful assembly are pre-emptively suppressed or responded to with violence or threats of violence. Social-media platforms in Cuba offer a limited, but still significant, channel for government criticism and social activism in the face of authorities that have restricted basic civil liberties and opportunities for offline civic mobilization.

While Meta said that it took several steps to mitigate risks to users during the July 2021 protests in Cuba, and again during mass protests planned for November 2021, it did not disclose any risk-mitigation measures at the time the case content was posted. To prepare for future occasions when calls for protests are expected to occur in places where protest will be responded to with violence or threats of violence from public authorities, and to ensure that such calls are reviewed and enforced accurately and with contextual nuance, Meta should consider how the political context could influence its policy and enforcement choices.

In order to address these concerns about moderating content that comes from closed civic spaces, the Board reiterates Recommendations #1 and #8 from the cross-check policy advisory opinion, noting their relevance to the Cuban context and content considered here. Recommendation #1 urged Meta to have a list-based over-enforcement prevention program to protect expression in line with Meta's human rights responsibilities. Over-enforcement prevention lists afford users that are included with additional opportunities for human review

of their posts that are initially identified as violating Meta's policies, with the aim of avoiding over-enforcement, or false positives. Recommendation no. 8 said that Meta should create such lists with local input. Meta has agreed to implement both recommendations in part, with implementation currently in progress.

9. Oversight Board Decision

The Oversight Board overturns Meta's decision to remove the post.

10. Recommendations

The Oversight Board decided not to issue new recommendations in this decision given the relevance of previous recommendations issued in other cases. The Board is aware of the cross-check status of the content creator's account at the time the content was reviewed and removed. Nevertheless, the Board still found recommendations no. 1 and no. 8 from the cross-check policy advisory opinion, in which the Board provides Meta with guidance for putting cross-check lists together, to be of great importance in this case given the context in Cuba. The Board believes that Meta should follow that guidance closely so that other accounts sharing valuable political speech, like the one in this case, are added to the list in order to benefit from additional layers of content review. For accounts already included in the list, the Board highlights the importance of recommendation no. 3 from the cross-check policy advisory opinion, which aims to improve the accuracy of enhanced content review for accounts on the list. Extending the opportunity for additional layers of content review, and the possibility of contextual information being incorporated in content moderation decisions, to more accounts that merit inclusion in the list – from a human rights perspective – is especially important in closed civic spaces, such as the one considered in this case.

- Recommendation no. 1, which urged Meta to have a list-based over-enforcement prevention program to protect expression in line with Meta's human rights responsibilities. This system is to be distinct from the system that protects expression that Meta views as a business priority, and should make sure Meta provides additional layers of review to content posted by, among others, human rights defenders.
- Recommendation no. 8, which said that Meta should use specialized staff, with the benefit of local input, to create over-enforcement prevention lists.

- Recommendation no. 3, which called on Meta to improve how its workflow dedicated to meet Meta's human rights responsibilities incorporates context and language expertise on enhanced review, specifically at decision making levels.

The Oversight Board further reiterates guidance provided to Meta throughout this and previous decisions to make sure context is appropriately factored into content moderation decisions and policies are sufficiently clear, to both users and content reviewers (Violence against women cases). This includes updating internal guidance provided to content reviewers where relevant in order for the company to address any lack of clarity, gaps or inconsistencies which may result in enforcement errors, such as the one in this case.

- Recommendation no. 2 from the Violence against women cases, which urged Meta to update guidance to its at-scale moderators with specific attention to rules around qualification, since the current guidance makes it virtually impossible for moderators to make the correct decisions.

***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. 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.