OVERTURNED

2022-005-FB-UA

Mention of the Taliban in news reporting

The Oversight Board has overturned Meta's original decision to remove a Facebook post from a news outlet page reporting a positive announcement from the Taliban regime in Afghanistan on women and girls' education.

Policies and topics

- Journalism, News events, Politics
- Dangerous individuals and organizations

Region and countries

- Central and South Asia
- Afghanistan

Platform

- Facebook

Attachments

Pashto translation

Dari translation

Mention of the Taliban in news reporting public comments

This decision is also available in Urdu (via the 'language' tab accessed through the menu at the top of this screen), in Pashto (here), and in Dari (here).

په پښتو ژبه د دغې پريکړې د لوستلو لپاره دلته کلیک وکړئ.

برای خواندن این تصمیم اینجا فشار دهید.

Case summary

The Oversight Board has overturned Meta's original decision to remove a Facebook post from a news outlet page reporting a positive announcement from the Taliban regime in Afghanistan on women and girls' education. Removing the post was inconsistent with Facebook's Dangerous Individuals and Organizations Community Standard, which permits reporting on terrorist groups, and Meta's human rights

responsibilities. The Board found Meta should better protect users' freedom of expression when it comes to reporting on terrorist regimes and makes policy recommendations to help achieve this.

About the case

In January 2022, a popular Urdu-language newspaper based in India posted on its Facebook page. The post reported that Zabiullah Mujahid, a member of the Taliban regime in Afghanistan and its official central spokesperson, had announced that schools and colleges for women and girls would reopen in March 2022. The post linked to an article on the newspaper's website and was viewed around 300 times.

Meta found that the post violated the Dangerous Individuals and Organizations policy which prohibits "praise" of entities deemed to "engage in serious offline harms," including terrorist organizations. Meta removed the post, imposed "strikes" against the page administrator who had posted the content and limited their access to certain Facebook features (such as going live on Facebook).

The user appealed and after a second human reviewer assessed the post as violating, it was placed in a queue for the High-Impact False Positive Override (HIPO) system. HIPO is a system Meta uses to identify cases where it has acted incorrectly, for example, by wrongly removing content. However, as there were less then 50 Urdu-speaking reviewers allocated to HIPO at the time, and the post was not deemed high priority, it was never reviewed in the HIPO system.

After the Board selected the case, Meta decided the post should not have been removed as its rules allow "reporting on" terrorist organizations. It restored the content, reversed the strike, and removed the restrictions on the user's account.

Key findings

The Oversight Board finds that removing this post is not in line with Facebook's Dangerous Individuals and Organizations Community Standard, Meta's values, or the company's human rights responsibilities.

The Dangerous Individuals and Organizations Community Standard prohibits "praise" of certain entities, including terrorist organizations. "Praise" is defined broadly in both the Community Standard, and the internal guidance for moderators. As a result, the Board understands why two reviewers interpreted the content as praise. However, the Community Standard permits content that "reports on" dangerous organizations. The Board finds this allowance applies in this case.

The Board also finds that removing the post is inconsistent with Meta's human rights responsibilities; it unjustifiably restricts freedom of expression, which encompasses the right to impart and receive information, including on terrorist groups. This is particularly important in times of conflict and crisis, including where terrorist groups exercise control of a country.

The Board is concerned that Meta's systems and policies interfere with freedom of expression when it comes to reporting on terrorist regimes. The company's Community Standards and internal guidance for moderators are not clear on how the praise prohibition and reporting allowance apply, or the relationship between them. The fact that two reviewers found the post was violating suggests that these points are not well understood. The Board is concerned that Meta's default is to remove content under the Dangerous Individuals and Organizations policy if users have not made it clear that their intention is to "report." The Board is also concerned that the content was not reviewed within the HIPO system.

This case may indicate a wider problem. The Board has considered a number of complaints on errors in enforcing the Dangerous Individuals and Organizations policy, particularly in languages other than English. This raises serious concerns, especially for journalists and human rights defenders. In addition, sanctions for breaching the policy are unclear and severe.

The Oversight Board's decision

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

The Board recommends that Meta:

- Investigate why changes to the Dangerous Individuals and Organizations policy were not translated within the target timeframe, and prevent such delays being repeated.
- Make the public explanation of its "strike" system more comprehensive and accessible.

- Narrow the definition of "praise" in the Known Questions (internal guidance for moderators) by removing the example of content that "seeks to make others think more positively about" dangerous organizations.
- Revise its Implementation Standards (internal guidance for moderators) to clarify that the reporting allowance in the Dangerous Individuals Organizations policy permits positive statements. The Known Questions should clarify the importance of protecting news reporting in conflict or crisis situations.
- Assess the accuracy with which reviewers enforce the reporting allowance to the Dangerous Individuals and Organizations policy to identify the cause of errors.
- Conduct a review of the HIPO system to examine whether it can more effectively prioritize potential errors in enforcing exceptions to the Dangerous Individuals and Organizations policy.
- Increase the capacity allocated to HIPO review across all languages.

*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 original decision to remove a Facebook post on the page of a popular Urdu language newspaper in India. This post reports an announcement from a prominent member and spokesperson of the Taliban regime in Afghanistan regarding women and girls' education in Afghanistan. Meta reversed its decision as a result of the Board selecting this case, and reversed sanctions on the administrator's account. The Board finds that the post does not violate the Dangerous Individuals and Organizations Community Standard because the policy allows "reporting on" designated entities. The Board is concerned by Meta's broad definition of "praise," and the lack of clarity to reviewers on how to enforce exceptions to the policy on news reporting of actions taken by designated entities that are exercising control of a country. This interferes with the ability of news outlets to report on the actions and statements of designated entities in situations like this one, where the Taliban regime forcibly removed the recognized government of Afghanistan. The Board finds Meta did not meet its responsibilities to prevent or mitigate errors when enforcing these policy exceptions. The decision recommends that Meta change its policy and enforcement processes on the Dangerous Individuals and Organizations Community Standard.

2. Case description and background

In January 2022, the Facebook page of a news outlet based in India shared a text post in Urdu containing a link to an article on its own website. Meta states that the post was viewed about 300 times. The post reported that Zabiullah Mujahid, acting as "Culture and Information Minister" and official central spokesman for the Taliban regime in Afghanistan, had announced that schools and colleges for girls and women would open at the start of the Afghan New Year on March 21. The linked article contains a fuller report on the announcement.

The news outlet is an Urdu-language newspaper based in Hyderabad, India, a city with a high number of Urdu-speaking residents. It is the largest-circulated Urdu newspaper in the country and claims a daily readership of more than a million people. There are approximately 230 million Urdu-speakers around the world.

No state has given formal diplomatic recognition to the Taliban regime in Afghanistan since the group seized power in August 2021. Schools and colleges for girls and women did not open at the start of the Afghan New Year as the spokesperson announced they would, and girls aged 12 and older (from the sixth grade on) and women remain barred from attending school at the time of the Board's decision in this case.

On January 20, 2022, a Facebook user clicked "report post" on the content but did not complete their complaint. This triggered a classifier (a machine learning tool trained to identify breaches of Meta's Community Standards) that assessed the content as potentially violating the Dangerous Individuals and Organizations policy and sent it for human review. An Urdu-speaking reviewer determined that the content violated the Dangerous Individuals and Organizations policy and removed it on the same day it was posted. Meta explained this was because it "praised" a designated organization. The Taliban is a Tier 1 designated terrorist organization under Meta's Dangerous Individuals and Organizations policy. As a result of the violation, Meta also applied both a severe strike and a standard strike against the page administrator. In general, while content posted to Facebook pages appears to come from the page itself (for example, the news outlet), they are authored by page administrators with personal Facebook accounts. Strikes result in Meta imposing temporary restrictions on users' ability to perform essential functions on the platform (such as share content), known as "feature-limits," or disabling of the account. Severe strikes result in stronger penalties. In this case, the strikes meant that a three-day and an additional, longer feature-limit was imposed against the page's administrator. The former prevented the user from creating new public content and creating or joining Messenger rooms. The latter prevented

the user going live on Facebook, using ad products, and creating or joining Messenger rooms. Additionally, the news outlet page itself also received one standard strike and one severe strike.

On January 21, the administrator of the news outlet page ("the user") appealed the removal of the content to Meta. The content was reviewed by another Urdu-speaking reviewer who also found that the content violated the Community Standard on Dangerous Individuals and Organizations. Though the content was then placed in a queue for identifying and reversing "false positive" mistakes (content wrongly actioned for violating the Community Standards), known as High Impact False Positive Override (HIPO), it received no additional review. According to Meta, this was because of the number of Urdu speaking HIPO reviewers in mid-2022 and because the content in this case, after it was removed, was not given a priority score by Meta's automated systems as high as other content in the HIPO queue at that time.

As a result of the Board's selection of the user's appeal for review, Meta determined that its original removal decision was in error because its Community Standards allow "reporting on" designated organizations and individuals. On February 25, 2022, Meta subsequently restored the content. Meta also removed the longer feature-limit it had imposed and reversed the strikes against the administrator's account and the page.

3. Oversight Board authority and scope

The Board has authority to review Meta's decision following an appeal from the user 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 policy advisory statements with non-binding recommendations that Meta must respond to (Charter Article 3, Section 4; Article 4).

4. Sources of authority

The Oversight Board considered the following authorities and standards:

I. Oversight Board decisions:

The most relevant prior Oversight Board decisions include:

- "Shared Al Jazeera post " (Case decision 2021-009-FB-UA): The Board recommended that Meta provide public criteria and examples in its Dangerous Individuals and Organizations Community Standard for the following allowances to the policy: "neutral discussion"; "reporting"; and "condemnation."
- "Öcalan's isolation" (Case decision 2021-006-IG-UA): The Board recommended that Meta clarify in its public-facing Dangerous Individuals and Organizations Community Standards how users can make their intent clear when posting. It reiterated recommendations that Meta disclose either the full list of designated individuals and organizations or an illustrative list. The Board also called for enhanced transparency reporting on error rates for enforcing its prohibition on praise and support of designated individuals and organizations, broken down by region and language.
- "Punjabi Concern over the RSS in India" (Case decision 2021-003-FB-UA): The Board recommended that Meta should aim to make its Community Standards accessible in all languages widely spoken by its users. The Board also expressed concerns that Meta's rules on restrictions are spread across many locations and not all found in the Community Standards, as one would expect.
- "Nazi quote" (Case decision 2020-005-FB-UA): The Board recommended that Meta provide examples of "praise," "support" and "representation" in the Community Standard on Dangerous Individuals and Organizations.

II. Meta's content policies:

The Community Standard on Dangerous Individuals and Organizations states that Facebook does "not allow organizations or individuals that proclaim a violent mission or are engaged in violence to have a presence on Facebook."

Meta divides its designations of "dangerous" entities into three tiers, explaining these "indicate the level of content enforcement, with Tier 1 resulting in the most extensive enforcement because Meta states that these entities have the most direct ties to offline harm." Tier 1 designations are focused on "entities that engage in serious offline harms," including "terrorist, hate and criminal organizations." Meta removes "praise," "substantive support," and "representation" of Tier 1 entities as well as their leaders, founders, or prominent members. Meta designates the Taliban as a Tier 1 entity.

The Community Standards define “praise” as any of the following: “speak positively about a designated entity or event”; “give a designated entity or event a sense of achievement”; “legitimiz[e] the cause of a designated entity by making claims that their hateful, violent, or criminal conduct is legally, morally, or otherwise justified or acceptable”; or “align[...] oneself ideologically with a designated entity or event.”

Meta recognizes that “users may share content that includes references to designated dangerous organizations and individuals to report on, condemn, or neutrally discuss them or their activities.” Meta says its policies are designed to “allow room for these types of discussions while simultaneously limiting risks of potential offline harm.” However, Meta requires “people to clearly indicate their intent when creating or sharing such content. If a user's intention is ambiguous or unclear, we default to removing content.”

III. Meta’s values:

The value of "Voice" is described as "paramount":

The goal of our Community Standards is to create a place for expression and give people a voice. Meta wants people to be able to talk openly about the issues that matter to them, even if some may disagree or find them objectionable.

Facebook limits "Voice" in the service of four values. "Safety" is the most relevant in this case:

We’re committed to making Facebook a safe place. We remove content that could contribute to a risk of harm to the physical security of persons. Content that threatens people has the potential to intimidate, exclude or silence others and isn’t allowed on Facebook.

IV. International human rights standards:

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, where 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 human rights standards:

- The right to freedom of opinion and expression: Article 19, International Covenant on Civil and Political Rights (ICCPR); Human Rights Committee, General Comment No. 34, (2011); Human Rights Council, Resolution on the Safety of Journalists, A/HRC/RES/45/18, 2020; UNESCO Brisbane Declaration on Freedom of Information: The right to know; UN Special Rapporteur on freedom of opinion and expression, A/74/486, 2019.
- The right to education: Article 13, International Covenant on Economic Social and Cultural Rights; Article 10, Convention on the Elimination of all forms of Discrimination against Women; Articles 28-29 on the Convention of the Rights to the Child; the UNESCO Convention against Discrimination in Education, 1960.
- The right to non-discrimination: ICCPR Articles 2 and 26.
- The right to life: ICCPR Article 6; Human Rights Committee, General Comment No. 36, 2018.
- The right to security of person: ICCPR Article 9, as interpreted by General Comment No. 35, para. 9, Human Rights Committee, 2014.

5. User submissions

In their statement to the Board, the user states that they are a representative of a media organization and do not support extremism. The user says that their articles are based on national and international media sources and that this content was shared to provide information about women and girls’ education in Afghanistan. Also, the user says they always ensure the content they share is in the public interest and that it is acceptable under Meta’s Community Standards.

6. Meta’s submissions

Upon re-examining its original decision, Meta decided that the content in this case should not have been removed as praise of a designated organization under the Dangerous Individuals and Organization policy. Meta explained that the underlying news context meant the content should have benefitted from the policy allowance for users to report on designated entities.

Meta explained that the post and linked article include reporting on school reopening dates and details, an issue of public interest. According to Meta, the Dangerous Individuals and Organization policy allows news reporting that mentions designated entities. To benefit from the

allowances built into the Dangerous Individuals and Organization policy, Meta clarified that "we require people to clearly indicate their intent. If a user's intention is ambiguous or unclear, we default to removing content." Meta also explained that it prefers its reviewers not to infer intent as this "helps to reduce subjectivity, bias, and inequitable enforcement during content review while maintaining the scalability of our policies."

Meta informed the Board that it was unable to explain why two human reviewers incorrectly removed the content and did not properly apply the allowance for reporting. The company noted that moderators are not required to document the reasons for their decision beyond classifying the content as part of their review — in this case, as a violation of Meta's Dangerous Individuals and Organizations policy on the grounds of praise.

In response to the Board questioning whether praise of dangerous organizations can be disseminated as part of news reporting, Meta stated its policy "allows news reporting where a person or persons may praise a designated dangerous individual or entity."

Answering the Board's question on the difference between standard and severe strikes, Meta explained that the strike system contains two tracks for Community Standards enforcement: one that applies to all violation types (standard), and one that applies to the most egregious violations (severe). Meta states that all violations of the Dangerous Individuals and Organizations policy are treated as severe. The company explained to the Board that severe strikes are those that apply stricter penalties against more serious harms, and limit access to high-risk services such as Facebook Live Video and ads. Meta also referenced a page in its Transparency Centre on "Restricting Accounts" (updated February 11, 2022) that it says explains its approach to strikes.

In response to the Board's questions, Meta provided further explanation of its systems for correcting enforcement errors and how those impacted this case, leading to the Board asking several follow-up questions. The content in this case was automatically detected and sent to a High Impact False Positive Override (called HIPO by Meta) channel. This is a system designed to correct potential false positive mistakes *after* action is taken on the content. Meta clarified to the Board that this system contrasts with Meta's General Secondary Review system (part of the cross-check program), which is designed to *prevent* false positive mistakes *before* action is taken on the content. Content sent to the HIPO channel joins a queue for additional review, but review will only occur where capacity allows. The position of content in the queue depends on a priority score automatically assigned to the content. Meta explained that content is prioritized for HIPO review based on factors including: topic sensitivity (if a topic is trending or sensitive); false positive probability; predicted reach (the estimated number of views the content might obtain); and entity sensitivity (the identity of the group or user sharing the content). Meta explained that content can be restored in two ways: either the specific piece of content is reviewed by moderators and found to be non-violating; or Meta's automated systems find that the content matches other content that has been reviewed and determined to be non-violating.

The page of the news outlet was previously subject to cross-check, but as part of a platform wide update to the cross-check system, the page was not subject to cross-check when the case content was reviewed, and cross-check did not impact the review of the case content. According to Meta, cross-check now includes two systems: General Secondary Review applies to all organic content on Facebook and Instagram; Early Response Secondary Review applies to all content posted by specific listed entities, including some news outlets. Meta stated that when content from those specific entities is identified as violating a content policy, instead of being enforced, it is sent for additional review. It is first sent to Meta's Markets team. If a reviewer on this team finds the content is not violating, the process ends, and the content remains on the platform. However, if a reviewer on this team finds the content is violating, it is escalated to another team. This team, the Early Response team, is made up of specialized Meta content reviewers. A reviewer on this team would need to find the content violating before it can be removed.

At the time the case content was identified as violating, the news outlet page was not on the Early Response Secondary Review list in the current cross-check system. Additionally, the case content in question was not reviewed as part of the General Secondary Review system, which would also involve additional review before enforcement. According to Meta, the content was sent to the HIPO channel after it was removed, but it was not prioritized for human review. It did not receive an additional human review "due to the capacity allocated to the market" and because the content in this case was not given a priority score by Meta's automated systems as high as other content in the HIPO queue at that time. Content prioritized by HIPO is only reviewed by outsourced reviewers *after* an enforcement action is taken. Meta allocates Urdu reviewers to different workflows based on need. These reviewers are shared across multiple review types, meaning they are not solely dedicated to a single workflow. In mid-2022, Meta's HIPO workflow had less than 50 Urdu reviewers based on that need at the time.

The Board asked Meta 34 questions. Meta responded to 30 fully, three partially and declined to answer one. The partial responses were to questions on: providing the percentage of removals under the Dangerous Individuals and Organizations policy that are restored on appeal or second review; accuracy rates for enforcing the prohibitions on praise and support in at-scale review; and how Meta determines intent for the reporting allowance and applicable contextual factors. Meta left one of the questions on providing data regarding the volume of Dangerous Individuals and Organizations content that is removed through automation versus human review unanswered on the grounds that it was unable to verify the requested data in the time available.

to verify the requested data in the time available.

7. Public comments

The Oversight Board received and considered six public comments related to this case. One of the comments was submitted from Asia Pacific and Oceania, four were from Europe, and one was from the United States and Canada.

The submissions covered the importance of access to social media to people who live in or near Afghanistan, concerns about limitations on the discussion of designated groups, and the public interest in allowing a wider range of media reporting on the Taliban's actions. Several public comments argued that the reliance of Meta's Dangerous Individuals and Organizations policy on the vague terms praise and support may suppress critical political discussion and disproportionately affect minority communities and the Global South. Public comments also criticized Meta for using US law as an "excuse" to prohibit "praise" of designated groups, rather than be transparent that it is Meta's policy choice to restrict more expression than US law requires.

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

8. Oversight Board analysis

This case is significant because it shows that a lack of clarity in the definition of praise seems to be resulting in uncertainty among reviewers and users. It also considers important issues of content moderation as it applies to gender discrimination and conflict. This case is difficult because there is an interest in ensuring that terrorist groups or their supporters do not use platforms for propaganda and recruitment efforts. However, this interest, when applied too broadly can lead to censorship of any content that reports on these groups. The Board looked at the question of whether this content should be restored, and the broader implications for Meta's approach to content moderation, through three lenses: Meta's content policies, the company's values and its human rights responsibilities.

8.1 Compliance with Meta's content policies

I. Content rules

The Board finds that the content falls within the allowance that "users may share content that includes references to designated dangerous organizations (...) to report on (...) them or their activities" and is therefore not violating. The content is not violating despite the broad definition of praise provided in the public facing Community Standards, even as the content can be understood as speaking positively about an action of a designated entity, the Taliban, and may give them "a sense of achievement." The Board notes the Known Questions guidance provided to moderators for interpreting praise is even broader. It instructs reviewers to remove content for praise if it "makes people think more positively about" a designated group, making the meaning of "praise" less about the intent of the speaker but the effects on the audience. Reporting on a designated entity's claimed intentions to allow women and girls access to education, however dubious those claims, would arguably make people think more positively about that group. Given the instructions they were provided, the Board understands why two human reviewers (in error) would interpret the content as praise.

The Board accepts that Meta intends for its Community Standard on Dangerous Individuals and Organizations to make room for reporting on entities Meta has designated as dangerous, even if that reporting also meets the company's definition of praise. However, the Board does not think the language of the Community Standard or the Known Questions makes that definition clear. As a matter of fact, without specification, "praise" remains overbroad. The Community Standards do not provide any examples of what would constitute acceptable reporting. There is also no internal guidance to moderators on how to interpret this allowance.

II. Enforcement action

The Board notes that the moderation action in this case occurred after a user began to report the content but never finished the report. An automated system is triggered when this process is initiated, even if the user does not end up completing the report, and so the content was enqueued for human review. The reporting user was not made aware that their action could trigger consequences even if they decide against finishing the report, whereas a user would be told of the consequences of their report if they submitted it. Meta argues in its responses that the "automated report is not tied to the [reporting] user" but the Board finds this noteworthy when the whole process for this case began with a user initiating a report. Also, the Board is concerned that the reporting button does not provide users with sufficient information on the consequences of clicking it.

The enforcement actions taken in this case (the content removal, strikes, and feature-limitations), should not have been imposed as there was no underlying violation of the Community Standards. The Board is concerned that Meta's systems for preventing enforcement errors of this kind were ineffective, particularly given the severity of the sanctions imposed.

The Board notes that in this case, the page of the news outlet was previously subject to cross-check, but as part of a platform wide update to the cross-check system, the page was not subject to cross-check when the case content was reviewed, and cross-check did not impact the review of the case content. In Meta's current cross-check system, guaranteed secondary human review is provided to users on the Early Response Secondary Review List. While some news outlets' Facebook pages are on that list, this page is not. Being on an Early Response Secondary Review list also guarantees that Meta employees, and not at-scale reviewers, review the content before it can be removed. The Board finds it unlikely that this content would have been removed if the page were on the Early Response Secondary Review list at the time.

The Board commends Meta for the introduction of its HIPO system but is concerned that it did not lead to secondary review of a post that conformed with Meta's Community Standards. The content in this case did not receive an additional human review "due to the capacity allocated to the market" and because it was not given a priority score by Meta's automated systems as high as other content in the HIPO queue at that time. Given the public interest nature of the reporting in this case, and the identity of the page as a news outlet posting the content, it should have scored highly enough for additional review to have taken place. As explained by Meta, the factor "entity sensitivity" takes the identity of the posting entity into account and can lead to a higher ranking for content from news outlets, especially those reporting on significant world events. For the same reasons, the Board is concerned that the Urdu language queue only had less than 50 reviewers in mid-2022. The Board considers the size of the India market, the number of groups Meta has designated as dangerous in that region, and therefore the heightened importance of independent voices, warrant greater investment from the company in correcting (and ideally preventing) errors from occurring on such important issues.

8.2 Compliance with Meta's values

Content containing praise of dangerous groups may threaten the value of "Safety" for Meta's users and others because of its links to offline violence and its potential to "intimidate, exclude or silence others." However, there is no significant safety issue in this case as the content only reports on the announcement of a designated organization. "Voice" is particularly important in relation to media outlets as they provide their audiences with essential information and play a crucial role in holding governments to account. Removing the content in this case did not materially contribute to "Safety" and was an unnecessary restriction of "Voice."

8.3 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, and that Meta should have more effective systems in place for preventing and correcting such errors. Meta adhering to its human rights responsibilities is particularly important in the context of crisis or conflict situations. Following the forceful takeover of a government by a group renowned for human rights abuses and due to the importance of informing the public of the acts of such designated groups, the company should be particularly attentive to protecting news reporting about that group.

Freedom of expression (Article 19 ICCPR)

Article 19 of the ICCPR provides protection of the right to freedom of expression and encompasses the right of all individuals to impart information and to receive it. International human rights law places particular value on the role of journalism in providing information that is of interest to the public. The UN Human Rights Committee has stated that "a free, uncensored and unhindered press or other media is essential in any society to ensure freedom of opinion and expression and the enjoyment of other Covenant rights" (General Comment No. 34, at para. 13). Social media platforms like Facebook have become a vehicle for transmitting journalist's reporting around the world, and Meta has recognized its responsibilities to journalists and human rights defenders in its corporate human rights policy.

The right to freedom of expression encompasses the ability of Meta's users to access information about events of public interest in Afghanistan, especially when a designated dangerous group forcibly removed the recognized government. It is imperative that users, including commentators on Afghanistan within and outside the country, and the general public have access to real-time reporting on the situation there. The Taliban's approach to media freedom in the country makes the role of international reporting even more important. The information in this case would be essential to people concerned about girls' and women's equal right to access education. This remains the case even when the Taliban fails to meet those commitments.

Article 19, para. 3 of the ICCPR, as interpreted by the Human Rights Committee in General Comment No. 34, requires that where restrictions on

expression are imposed by a state, they must meet the requirements of legality, legitimate aim, and necessity and proportionality. The Board applies these international standards to assess whether Meta complied with its human rights responsibilities.

I. Legality (clarity and accessibility of the rules)

The principle of legality requires laws that States use to limit expression to be clear and accessible, so people understand what is permitted and what is not. Further, it requires laws restricting expression to be specific, to ensure that those charged with their enforcement are not given unfettered discretion (General Comment 34, para. 25). The Human Rights Committee has warned that "offences of 'praising', 'glorifying', or 'justifying' terrorism, should be clearly defined to ensure that they do not lead to unnecessary or disproportionate interference with freedom of expression. Excessive restrictions on access to information must also be avoided" (General Comment No. 34, at para. 46; see also: UN Special Rapporteur on counter-terrorism and human rights at paras 36-37(Report A/HRC/40/52)).

Following its approach in previous cases, the Board applies these principles to Meta's content rules. While the Board welcomes that Meta's policies on Dangerous Individuals and Organizations contain more detail now than when the Board issued its first recommendations in this area, serious concerns remain.

For users reporting on the Taliban, it is unclear whether the Taliban remains a designated dangerous entity when it forcibly removed the recognized government of Afghanistan. The Board has previously recommended that Meta discloses either a full list of designated entities, or an illustrative one, to bring users clarity ("Nazi quote" case, "Ocalan's isolation" case). The Board regrets the lack of progress on this recommendation, and notes that while the company has not disclosed this information proactively, whistleblowers and journalists have sought to inform the public by disclosing a version of the "secret" list publicly.

As noted previously in this decision (see section 8.1), the definition of "praise" in the public-facing Community Standards as "speaking positively about" a designated entity is too broad. For people engaged in news reporting, it is unclear how this rule relates to the reporting allowance built into the same policy. According to Meta, this allowance permits news reporting even where a user praises the designated entity in the same post. The Board finds the relationship between the "reporting" allowance in the Dangerous Individuals and Organizations policy and the overarching newsworthiness allowance remains unclear to users. In the "Shared Al Jazeera post" case, the Board recommended that Meta provides criteria and illustrative examples in the Community Standards on what constitutes news reporting. Meta responded in the Q1 2022 implementation report that it was currently consulting with several teams internally to develop criteria to help users understand what constitutes news reporting. It said it expects to conclude this process by Q4 2022.

The Board remains concerned that changes to the relevant Community Standard are not translated into all available languages and there are inconsistencies across languages. Following the Board's "Shared Al Jazeera post" decision, the US English version of the Dangerous Individuals and Organizations policy was amended in December 2021 to change the discretionary "we may remove content" to "we default to remove content" when a user's intentions were unclear. However, other language versions of the Community Standards, including in Urdu and UK English, do not reflect this change. While the company has stated publicly in response to previous recommendations from the Board (Meta Q4 2021 Quarterly Update on the Oversight Board) that it aims to complete translations into all available languages in four to six weeks, it appears the relevant policy line for this case has not been completed after five months. Therefore, the policy is not equally accessible to all users, making it difficult for them to understand what is permitted and what is not.

The Board is also concerned that Meta has not done enough to clarify to its users how the strikes system works. While a page on "Restricting Accounts" in Meta's Transparency Centre contains some detail, it does not comprehensively list the feature-limits the company may apply and their duration. Nor does it list the "set periods of time" for severe strikes as it does for standard strikes. This is especially concerning because severe strikes carry more significant penalties and there is no mechanism in place for appealing account-level sanctions separately from the content decision. Even when the content is restored, feature-limits cannot always be fully reversed. In this case, for instance, the user had already experienced several days of feature-limits which were not fully rectified when Meta reversed its decision.

That two Urdu-speakers assessed this content as violating further indicates that the praise prohibition, and its relationship to the reporting allowances, is unclear to those tasked with enforcing the rules. Content reviewers are provided with internal guidance on how to interpret the rules (the Known Questions and the Implementation Standards). The Known Questions document defines praise as content that "makes people think more positively about" a designated group. This is arguably broader than the public-facing definition in the Community Standards, making the meaning of "praise" less about the intent of the speaker than the effects on the audience. Also, neither the Community Standards, nor the Known Questions document constrain the reviewer's discretion on restricting freedom of speech. Standard dictionary definitions of "praise" are not this broad, and as phrased the rule captures statements of fact, including impartial journalistic statements, as

well as opinion. In response to the Board's questions, Meta clarified that the reporting allowance allows *anyone*, and not only journalists, to speak positively about a designated organization in the context of reporting. However, this clarity is not provided to reviewers in internal guidance. Meta admits this guidance does not provide reviewers with a definition of how to interpret "reporting on."

II. Legitimate aim

The Oversight Board has previously recognized that the Dangerous Individuals and Organizations policy pursues the aim of protecting the rights of others, including the right to life, security of person, and equality and non-discrimination (Article 19(3) ICCPR, Oversight Board decision "Punjabi Concern over the RSS in India"). The Board further recognises that propaganda from designated entities, including through proxies presenting themselves as independent media, may pose risks of harm to the rights of others. Seeking to mitigate those harms through this policy is a legitimate aim.

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). Meta has acknowledged that the removal of content in this case was not necessary, and therefore additional sanctions should not have been imposed on the user.

The Board understands that when moderating content at scale, mistakes will be made. However, the Board does receive complaints on errors in enforcing the Dangerous Individuals and Organizations policy that affect reporting, particularly in languages other than English, which raises serious concerns (see "Shared Al Jazeera post" decision, "Ocalan's isolation " decision). The UN Human Rights Committee has emphasized that "the media plays a crucial role in informing the public about acts of terrorism and its capacity to operate should not be unduly restricted. In this regard, journalists should not be penalized for carrying out their legitimate activities" (General Comment 34, para 46). Meta therefore has a responsibility to prevent and mitigate its platforms' negative human rights impact on news reporting.

The Board is concerned that the type of enforcement error in this case may be indicative of broader failures in this regard. Those engaged in regular commentary on the activities of Tier 1 dangerous individuals and organizations face heightened risks of enforcement errors leading to their accounts facing severe sanctions. This may undermine their livelihoods and deny the public access to information at key moments. The Board is concerned the policy of defaulting to remove content when the intent to report on dangerous entities is not clearly indicated by the user may be leading to over-removal of non-violating content, even where contextual cues make clear the post is, in fact, reporting. Moreover, the system for mistake prevention and correction did not benefit this user as it should have. This indicates problems with how the ranker within the HIPO system prioritized the content decision for additional review, which meant it never reached the front of the queue. It also raises questions about the resources allocated to human review of the HIPO queue potentially being insufficient for Urdu language content. In this case, the enforcement error and failure to correct it denied a number of Facebook users access to information on issues of global importance and hampered a news outlet in carrying out its journalistic function to inform the public.

Journalists may report on events in an impartial manner that avoids the kind of overt condemnation that reviewers may be looking to see. To avoid content removals and account sanctions, journalists may engage in self-censorship, and may even be incentivized to depart from their ethical professional responsibilities. Further, there have been reports of anti-Taliban Facebook users avoiding mentioning the Taliban in posts because they are concerned about being subjected to erroneous sanctions.

The Board also notes that Meta has issued what it calls "spirit of the policy" exceptions related to the Taliban. This indicates recognition from Meta that at times, its approach under the Dangerous Individuals and Organizations policy is producing results that are inconsistent with the policy's objectives, and therefore do not meet the requirement of necessity. Internal company materials obtained by journalists reveal that in September 2021, the company created an exception "to allow content shared by the [Afghanistan] Ministry of Interior" on matters such as new traffic regulations, and to allow two specific posts from the Ministry of Health in relation to COVID-19. Other exceptions have reportedly been more tailored and shorter-lived. For 12 days in August 2021, "government figures" could reportedly acknowledge the Taliban as the "official gov of Afghanistan [*sic*]" without risking account sanctions. From late August 2021 to September 3, users could "post the Taliban's public statements without having to 'neutrally discuss, report on, or condemn' these statements." Meta spokespersons acknowledged that some ad hoc exceptions were issued. In a Policy Forum on Crisis Policy Protocol, on January 25, 2022, Meta stated that it will deploy "policy levers" in crisis situations and provided the example of allowing "praise of a specific designated org (e.g. a guerrilla group signing a peace treaty)." These exceptions to the general prohibition on praise could cause more uncertainty for reviewers, as well as for users who may not be aware if or when an exception applies. They show that there are situations when Meta has reportedly recognized a more nuanced approach to content is warranted dealing with a designated entity that overthrows a legitimate government and assumes territorial control.

The Board finds that removing the content from the platform was an unnecessary and disproportionate measure. The volume of these enforcement errors, their effects on journalistic activity, and the failure of Meta's error-prevention systems have all contributed to this conclusion.

9. Oversight Board decision

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

10. Policy advisory statement

Content policy

1. Meta should investigate why the December 2021 changes to the Dangerous Individuals and Organizations policy were not updated within the target time of six weeks, and ensure such delays or omissions are not repeated. The Board asks Meta to inform the Board within 60 days of the findings of its investigation, and the measures it has put in place to prevent translation delays in future.

2. Meta should make its public explanation of its two-track strikes system more comprehensive and accessible, especially for "severe strikes." It should include all policy violations that result in severe strikes, which account features can be limited as a result and specify applicable durations. Policies that result in severe strikes should also be clearly identified in the Community Standards, with a link to the "Restricting Accounts" explanation of the strikes system. The Board asks Meta to inform the Board within 60 days of the updated Transparency Center explanation of the strikes system, and the inclusion of the links to that explanation for all content policies that result in severe strikes.

Enforcement

3. Meta should narrow the definition of "praise" in the Known Questions guidance for reviewers, by removing the example of content that "seeks to make others think more positively about" a designated entity by attributing to them positive values or endorsing their actions. The Board asks Meta to provide the Board within 60 days with the full version of the updated Known Questions document for Dangerous Individuals and Organizations.

4. Meta should revise its internal Implementation Standards to make clear that the "reporting" allowance in the Dangerous Individuals Organizations policy allows for positive statements about designated entities as part of the reporting, and how to distinguish this from prohibited "praise." The Known Questions document should be expanded to make clear the importance of news reporting in situations of conflict or crisis and provide relevant examples, and that this may include positive statements about designated entities like the reporting on the Taliban in this case. The Board asks Meta to share the updated Implementation Standards with the Board within 60 days.

5. Meta should assess the accuracy of reviewers enforcing the reporting allowance under the Dangerous Individuals and Organizations policy in order to identify systemic issues causing enforcement errors. The Board asks Meta to inform the Board within 60 days of the detailed results of its review of this assessment, or accuracy assessments Meta already conducts for its Dangerous Individuals and Organizations policy, including how the results will inform improvements to enforcement operations, including for HIPO.

6. Meta should conduct a review of the HIPO ranker to examine if it can more effectively prioritize potential errors in the enforcement of allowances to the Dangerous Individuals and Organizations Policy. This should include examining whether the HIPO ranker needs to be more sensitive to news reporting content, where the likelihood of false-positive removals that impacts freedom of expression appears to be high. The Board asks Meta to inform the Board within 60 days of the results of its review and the improvements it will make to avoid errors of this kind in the future.

7. Meta should enhance the capacity allocated to HIPO review across languages to ensure that more content decisions that may be enforcement errors receive additional human review. The Board asks Meta to inform the Board within 60 days of the planned capacity enhancements.

***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.

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. 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.