Pro-Iranian Government Posts

2026-41-IG-UA, 2026-42-IG-UA

Summary

To prevent influence operations, including those by governments, Meta should improve sharing between Meta's specialized teams of indicators of such operations, the Board has said, in examining two posts showing support for Iran's regime during military conflagrations and anti-government protests in the country.

A new system for sharing signals of coordinated inauthentic behavior should be activated in situations involving countries where there is a particular risk of generating or suffering such foreign or domestic campaigns, to enable teams to warn about potential organized influence operations and allow proactive reporting of patterns for review at scale. These influence operations could be perpetrated by governments or other organizations seeking to manipulate users.

The Board upholds Meta's decision to leave up one post but reverses the company's decision to leave up a second post.

Why This Matters

This case highlights the difficulty of identifying influence operations and their impact on users' free expression. It analyzes how Meta should address content that could be part of covert influence operations perpetrated on its platforms by governments or other organizations seeking to manipulate users. Such activities are designed to create the impression of a groundswell of support for a position that is in fact the product of an underhand campaign. Their impact is often difficult to measure, but they can erode trust in democratic systems, inflame social divisions and incite violence. However, while it is essential that platforms seek to counter the organized spread of

disinformation, care must be taken to avoid limiting free expression through overly zealous action.

About the Case

The first post was made amid diplomatic tensions and ahead of military strikes involving Iran, Israel and the United States, and the second post was made during widespread anti-government protests in Iran, which were followed by security forces' mass killings of demonstrators and an internet shutdown. The two recently made accounts that posted the content to Instagram include the same number in their usernames, which has been associated with a pro-Iranian regime cybercrime group.

The first post, made in June 2025, is an image of a man in military-style clothing with a caption in Farsi reading, "Just some patience Iran." Text over the image states: "The sound of the Iranian people's unity is making you tremble, isn't it? I swear by Imam Hussein, we will shock the world. We've been sleeping in our combat uniforms, we've been waiting for you." The content was viewed over 140,000 times. The account was later disabled.

In January 2026, another user posted a short video of what appears to be a motorcycle convoy of Iranian government security personnel driving through a city, with the caption in Farsi: "We are ready to give our lives for this country and its system until our very last breath," with related Farsi hashtags. The video is set to music with a voice singing a message similar to that of the caption. The content has been viewed over 9,000 times.

Multiple users reported the posts, and Meta determined they were not violative, leading to a user appealing both decisions. Neither appeal was reviewed, and the user appealed both cases to the Board. The appealing user said that the number in the accounts' usernames appeared to be a coded signal for coordinated accounts to recognize and amplify each other. They stated that the content went beyond ordinary political speech and instead was coded language and euphemism intended to intimidate and incite.

Prior to this case being selected by the Board, Meta had disabled the account that posted the content in the first case under the Authentic Identity Representation policy due to the account misrepresenting its identity and not representing a genuine user. After the Board selected the cases, Meta found that the account should have been disabled for being closely linked to multiple other accounts that had themselves been independently disabled for violating Meta's policies, including those against deceptive conduct and scams, under the company's Account Integrity policy.

Key Findings

The Board agreed with Meta that neither post violates the Dangerous Organizations and Individuals policy. The first post also does not contain explicit or implicit threats or calls for violence and therefore does not violate the Violence and Incitement policy. However, according to a majority of the Board, the second post – viewed in the context of the then-ongoing protests, the atrocities in response and the Iranian security forces' history of grave human rights violations – does violate the Violence and Incitement policy as it meets Meta's definition of a veiled threat. It should have been removed. This removal is consistent with Meta's human rights responsibilities to protect the protesters' right to life, given the escalating state violence against protesters at the time of the post. No intervention short of removal would adequately mitigate the risks the post presents. A minority of the Board, however, finds that the second post does not contain a threat or call for violence that violates the Violence and Incitement policy. Removing this post, based solely on the context, without there being even an implicit threat in the content, in the view of the minority, is neither necessary nor proportionate.

The Board finds that the Inauthentic Behavior policy could be improved by providing descriptions of the categories of behavior that the policy prohibits as well as an explanation that the enforcement is network-based rather than content-based.

Having reviewed responses to the Board's questions to Meta, the Board is satisfied that, in theory, when the threshold for breaching the Inauthentic Behavior policy is met,

disablement of the accounts involved is likely to be necessary and proportionate, with no other less intrusive means available to protect the rights of others. However, that does not ensure the policy will always be applied in a proportionate way.

Meta did not find that either account in this case was part of a coordinated inauthentic behavior network. However, the account that posted the first piece of content was found to be both inauthentic and connected to other accounts operated by the same user that had been disabled for violating Meta's policies. This was only uncovered by Meta after the Board selected the case and asked the company why the account had been disabled. This suggests that, even though coordinated inauthentic behavior investigators are skilled specialists with sophisticated technology, there are potential gaps in how Meta discovers such content and sends it to those specialists for review. Some signals or patterns that humans can recognize may later be useful to authenticity-related investigations. The Board therefore proposes better communications channeling to surface potentially suspicious content.

Meta should consider regularly reporting the number of accounts taken down for spreading disinformation, including when those accounts are not part of a network, which would be useful to researchers and civil society.

The Oversight Board's Decision

The Board upholds Meta's decision to leave up the first post, but reverses Meta's decision to leave up the second post.

The Board recommends that Meta:

- Create a system for signal sharing between Meta's teams specializing in coordinated inauthentic behavior and at-scale moderators. This system should be activated for situations involving countries facing particular risk of being impacted by foreign or domestic influence operations; enable coordinated inauthentic behavior review teams to flag content signals reviewers can use to

escalate content that may be useful to their investigations; and allow for at-scale reviewers to proactively report patterns that may warrant more comprehensive review.

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

Full Case Decision

1. Case Description and Background

This case concerns how Meta should address content that could be part of covert influence operations on its platforms perpetrated by governments or other organizations seeking to manipulate users. Such activities are designed to create the impression of a groundswell of support for a particular position. The decision analyzes the challenge of identifying such underhand campaigns by assessing two pieces of content posted in support of the Iranian government and security forces during periods of national tension and crisis.

Accounts whose usernames include a number that has been associated with a pro-Iranian regime cybercrime group posted each piece of content. These accounts were also created less than two months before the posts were made. These signals, along with the nature of the content, suggested the posts could be part of an organized online influence operation.

The first post, made on Instagram on June 12, 2025, is an image of a man in military-style clothing with a caption in Farsi reading: “Just some patience Iran.” Text over the image states: “The sound of the Iranian people’s unity is making you tremble, isn’t it? I swear by Imam Hussein, we will shock the world. We’ve been sleeping in our combat uniforms, we’ve been waiting for you.” The content was viewed over 140,000 times, received about 5,000 likes and just under 1,000 comments before the account was later disabled.

The image was posted after weeks of escalating international diplomatic pressure on Iran, on the same day that the International Atomic Energy Agency declared the country to be non-compliant with its nuclear non-proliferation obligations. The day after the post was made, Israel launched a series of military strikes against Iran, killing several of the regime's senior military commanders and nuclear scientists. Tehran retaliated and the conflict escalated, with the U.S. commencing attacks on Iran on June 22, 2025. On June 24, 2025, a ceasefire announced by U.S. President Donald Trump took effect, but over the following months all parties resumed military action. Previous sanctions on the Iranian regime were reimposed, and the economic situation in the country worsened. The conflict led to the deaths of almost 1,200 Iranians, a third of whom were civilians, and 28 Israelis, all but one of whom were civilians.

Large-scale protests broke out in Iran on December 28, 2025, in response to the deteriorating situation in the country, including economic conditions. These protests quickly turned into mass nationwide demonstrations calling for an end to the Islamic Republic system and a transition to a new form of government. From the start of January 2026, security forces reportedly began killing and arresting protesters.

On January 6, 2026, another Instagram user posted a short video of what appears to be a motorcycle convoy of Iranian government security personnel driving through a city, with the caption in Farsi, "We are ready to give our lives for this country and its system until our very last breath," and hashtags including #Iran, #Special_Unit, #Protest and #Unrest. The video is set to music with a voice singing a message similar to that of the caption, claiming they will sacrifice their lives for Iran and the regime. The content has been viewed over 9,000 times, receiving more than 300 likes and fewer than 50 comments. On January 7 and 8, the protests peaked and the government cut off internet access in the country. Security forces carried out mass killings of protesters, firing on crowds and leaving thousands of people dead and many more injured. The protests continued until February 28, when the U.S. and Israel attacked Iran again.

Both posts were reported by multiple users and, after Meta's automated systems determined that they did not violate any of its policies, a single user appealed both

decisions. Neither appeal was prioritized for human review and, after they timed out, the user appealed both cases to the Board.

After the Board selected these cases, Meta confirmed that the individual posts did not violate its policies. The company then investigated the accounts and determined that they did not appear to be part of an organized operation. Nevertheless, the Board decided to proceed with the case to highlight the difficulty of identifying influence operations and other coordinated inauthentic content and its impact on users' free expression. When the Board selected the cases, Meta disclosed that the account that posted the content in the first case had recently been disabled under the company's Account Integrity policy, which had the impact of removing the post from the platform. As of the time of writing, the account that posted the motorcycle convoy video remains online.

2. User Submissions

The user who appealed both pieces of content to the Board stated that the number used in both account names appeared to be a coded signal for coordinated accounts to recognize and amplify each other. They stated that the content went beyond ordinary political speech and instead was coded language and euphemism intended to intimidate and incite. The user further explained that the language about readiness and loyalty was an implicit threat to dissenters and opponents of the Iranian government. They also linked both posts to risks of offline harm and called for more proactive context-aware enforcement against coordinated networks.

3. Meta's Policies and Submissions

I. Content Policies

Meta determined that the Dangerous Organizations and Individuals policy and the Violence and Incitement policy were potentially relevant to these cases. The company assessed each piece of content against them.

The Dangerous Organizations and Individuals policy prohibits designated "organizations or individuals that proclaim a violent mission or are engaged in violence" from having a presence on Meta's platforms. Meta told the Board that neither post involved a reference to a designated entity and therefore were non-violating. The first case includes references to Iran and its military forces but not to any specific organization or individual. The pictured man in military-style clothing is not wearing any insignia or symbols of a designated entity. The second case depicts men on police motorcycles wearing fatigues with the hashtag referencing #Special_Unit, which Meta believes refers to Iran's national police force, often called by its Farsi acronym, FARAJA, which is not designated under the policy. The content does not reference any other organization or any individuals.

Meta's Violence and Incitement policy aims to "prevent potential real-world harm that may be related to content on" its platforms by removing "language that threatens, incites or facilitates violence." Meta told the Board that, in its opinion, neither post contained a threat of violence, any incitement to violence or any facilitation of violence towards an identifiable target.

Meta assessed the first case as containing "hyperbolic nationalist rhetoric" but no description of any specific violent act, method of harm or intention to commit violence against a person or group. The company explained that the post's language was vague, not addressed at a target and appeared to be a statement of support for Iran made at a time of intense international pressure.

On the second post, Meta explained that it includes patriotic language but does not articulate an intention, aspiration or call to commit a specific violent act. The company noted that, even if the content could be construed as containing threatening language, it was not directed against any identifiable target. Meta determined that the hashtags referencing the protests around the time of the post do not constitute a threat of violence against protesters, or incitement or facilitation of violence towards them.

The Violence and Incitement policy notes, in a subsection of items requiring "additional information and/or context to enforce," that Meta removes "coded statements where the method of violence is not clearly articulated, but the threat is veiled or implicit, as

shown by the combination of both a threat signal and context signal." A threat signal can be a coded statement "shared in a retaliatory context" or one that "acts as a threatening call to action." Context signals include a confirmation from local experts, information that the statement could lead to imminent violence, or the target of a threat reporting the content.

II. Actor and Behavior Policies

Most of Meta's Community Standards apply to individual pieces of content. They provide frameworks for posts to be individually assessed against rules and, if they violate them, for those posts to be removed or otherwise enforced against, such as with a warning screen. These content-based policies include the Violence and Incitement policy, which Meta applied to the content in these cases.

Some of the company's other policies, however, apply more broadly – focusing on the person or actor behind the content in question or patterns of behavior across the company's platforms. This decision touches on three of these policies.

One of them, the Authentic Identity Representation policy, prohibits accounts on Facebook or Instagram that deliberately misrepresent their identity to mislead or deceive others, evade enforcement or violate other policies. This actor-based prohibition focuses on the identity of the person behind an account and the content it posts.

A second such policy, the Inauthentic Behavior Community Standard, concerns behavior across Meta's platforms. Its rationale explains that "inauthentic behavior refers to a variety of complex forms of deception, performed by a network of inauthentic assets controlled by the same individual or individuals, with the goal of deceiving Meta or our community or to evade enforcement under the Community Standards." In response to questions from the Board, Meta explained that it uses this policy to examine coordinated deceptive activity across connected accounts, targeting networks that intentionally misrepresent themselves and engage in deceptive practices.

State-sponsored influence operations fall into a section of this policy dealing with the most sophisticated form of inauthentic behavior – what Meta calls "Coordinated Inauthentic Behavior" – which has three characteristics:

(1) coordination across a network of accounts to engage in particularly sophisticated forms of inauthentic behavior

(2) the use of false identities as a central part of the operation

(3) the use of adversarial methods to evade detection or appear authentic.

Enforcement takes a network disruption approach, removing entire coordinated networks.

Meta focuses on identifying patterns of deceptive behavior across networks of users by looking at a wide range of behavioral signals. This holistic behavior-based approach includes examining details such as the accounts' use and creation, geographic information, timing and coordination between different accounts, and posting patterns that emerge across a suspected network.

The third policy is the Account Integrity Community Standard, which is also both actor- and behavior-based. It sets out situations in which Meta will restrict and disable accounts. These include severe or persistent violations of policies and the creation or use of accounts to get around the removal of others. The policy states that Meta may disable accounts demonstrating "close linkage with a network of accounts or other entities that violate or evade [its] policies," "coordination within a network of accounts or other entities that persistently or egregiously violate [its] policies" and "activity or behavior indicative of a clear violating purpose through a network of accounts."

These three policies differ in what they focus on and the nature of the harm caused by each type of violation. The Authentic Identity Representation policy addresses situations where a single account attempts to mislead other users about what and who it represents. The policy does not require any connection to other accounts or any coordinated activity.

The prohibition on coordinated inauthentic behavior under the Inauthentic Behavior policy deals with situations where multiple actors use multiple distinct accounts and other assets to coordinate deceptive and inauthentic activity as a network.

The Account Integrity policy network prohibitions are designed for situations where one actor operates multiple accounts to evade enforcement. It does not require that the actor misrepresent its identity on any individual account.

III. The First Account

Prior to this case being selected by the Board, Meta had already disabled the account that posted the content in the first case. Meta found that the account was misrepresenting its identity and did not represent a genuine user, so disabled it under the Authentic Identity Representation policy. However, after the Board inquired into the circumstances of the account disablement, Meta determined that the disablement should have been under the Account Integrity policy instead, because the account was closely linked to multiple other accounts that had themselves been independently disabled for violating Meta's policies, including those against deceptive conduct and scams.

While the account was linked to a network of other accounts, it was not linked to other users as part of a broader coordinated network or influence operation. Accordingly, Meta determined that the account had not violated the Inauthentic Behavior policy.

The Board asked questions about Meta's moderation and enforcement efforts in Iran, how Meta identifies and enforces against influence operations and other forms of inauthentic behavior, and how Meta's related policies interact and are enforced. Meta responded to all questions.

4. Public Comments

The Board received five public comments that met the terms for submission. Four of the comments were submitted from the U.S and Canada, and one was from the Middle East. To read public comments submitted with consent to publish, click here.

The comments covered topics including: the need to use broad behavior-based rather than content-specific approaches to deal with influence operations; how influence operations work, including those carried out by or on behalf of the Iranian state; and the need to ensure that efforts to counter influence operations leave legitimate expression, particularly political speech, unaffected.

5. Oversight Board Analysis

The Board assessed the two posts against Meta's relevant content policies as well as through the lens of Meta's human rights commitments. The human rights assessment is based on the three-part test of legality, legitimate aim, and necessity and proportionality. This test is described in more detail in section 5(III)-(V) below, where the Board has reviewed the account disablement in this case and Meta's approach to influence operations against the company's human rights commitments.

I. Content Decisions

The Board finds that neither post violates the Dangerous Organizations and Individuals policy. While the man depicted in the first post is in military-style clothing, there is no insignia, badge or other indication that he is a member or supporter of the Islamic Revolutionary Guard Corps (IRGC) or any other entity designated by Meta. The religious figure cited in the first post is unaffiliated with any designated entity. In the second post, the motorcycle riders lack any visible affiliation with a designated group. While both posts reference the country of Iran in a supportive and patriotic fashion, there is no indication that this carries over to support for the IRGC or another designated entity.

Nor does the first post violate the Violence and Incitement policy, as it does not contain explicit or implicit threats or calls for violence. The post was made during a time of

extreme regional tension and stated that Iran will shock "the world" but does not state how or against whom specifically. While subject to multiple interpretations, the most likely is that the Iranian nation is united and will fight back against foreign attack. The statement that "We've been sleeping in our combat uniforms, we've been waiting for you" could be interpreted as saying that the user (ostensibly aiming to speak on behalf of nationalistic Iranians) is prepared to retaliate against aggression, but it does not threaten violence or call for it to be carried out. The Board notes that in times of crisis and international tension, statements of both support for and opposition to states and their armed forces as well as associated calls for violence are likely to increase. Social media platforms must ensure they do not remove legitimate political speech where no policies are violated.

However, a majority of the Board finds that the second post violates the Violence and Incitement policy when assessed in the circumstances it was posted in and that it should have been removed. Viewed in the context of the then ongoing protests, the atrocities being carried out in response and the Iranian security forces' history of grave human rights violations, the post meets Meta's definition of a veiled threat, which categorizes potentially ambiguous posts as threats if they have both a "threat signal" and a "context signal" that together make up an implied or disguised threat (see Content Targeting Human Rights Defender in Peru).

- Threat signal: The post satisfies the "threat" signal requirement specifically as "a threatening call to action." The video shows security forces on motorcycles riding through an Iranian city with the caption, "We are ready to give our lives for this country and its system until our very last breath." Given the context at the time, this supportive language should be interpreted as a call for the security forces to fight to the death against the protesters rather than against any external force. This is reinforced by the post's use of the hashtags #protest and #unrest. This statement, combined with footage that serves to emphasize the power of the security forces, can and should be interpreted as a threatening call for violence against protesters.

- Context signal: The post was made after the Iranian security forces had started killing protesters. This context, combined with the Iranian security forces' well-documented history of violent responses to protest, including the killing and execution of demonstrators, is sufficient context to find that this content could lead to imminent violence. This position is supported, after the fact, by the large-scale massacres the security forces carried out in the days immediately following this post, where they killed thousands of civilians.

These are difficult assessments to make, requiring contextual information and regional knowledge in a fast-moving and unreliable information environment. Meta assesses veiled threats only on escalation to specialist internal reviewers. At-scale moderators are not empowered to identify content that violates the company's escalations-only policies. The Board has previously raised concerns over the distinction between threats that do and do not require context to enforce (see Content Targeting Human Rights Defender in Peru).

The majority finds that removal of the second piece of content from the platform, as required by a proper interpretation of Meta's content policies, is consistent with Meta's human rights responsibilities. The Board has previously found that the Violence and Incitement policy's prohibition on "threats of violence against various targets" and "coded statements where the method of violence is not clearly articulated, but the threat is veiled or implicit" are sufficiently clear and pursue a legitimate aim, meeting the first two requirements of the three-part test.

Given the escalating state violence against protesters at the time the post was made, the Board finds that removing the content is a necessary and proportionate limitation on expression in order to protect the protesters' right to life as well as their right to peaceful assembly (Article 21, International Covenant on Civil and Political Rights (ICCPR)). No intervention short of removal would adequately mitigate the risks the post presents. The two posts can be differentiated by who they appear to target. The first seems to address hostile foreign states in the build-up to international armed conflict.

The second post, through its hashtags and timing, targets a domestic population protesting against the forces depicted in the video.

A minority of the Board, however, finds that the second post does not contain a threat or call for violence that violates the Violence and Incitement policy. The statement that the user would willingly die for his country, Iran, combined with a depiction of what appear to be security personnel does not in any clear way rise to the level of a threat, explicit or coded, to an identifiable target. The inclusion of the hashtags #protest and #unrest in the caption cannot be read, in and of itself, as identifying the protesters as a target, given the absence of a threat directed towards them. Additionally, despite the context referred to above, there are no signs in the content that can make this post a possible cause for potential real-world harm against identifiable targets. Removing this post, based solely on the context, without there being even an implicit threat in the content, in the view of the minority, is neither necessary nor proportional. As such, the minority also finds that Meta keeping the post on its platform was consistent with both its policies and human rights commitments.

Pro-government speech remains protected political speech even if the government is internationally condemned. However, no one, including governments, their agents and supporters, has a free speech right to threaten violence against lawful protesters.

II. Influence Operations and Account Decisions

Influence operations on social media are difficult to identify solely by examining individual posts or even individual accounts. As they often engage divisive social and political topics, it is difficult to distinguish between a user expressing their beliefs while unintentionally spreading misinformation and someone who is part of a concerted attempt to spread disinformation to achieve a specific goal. The impact of influence operations is often difficult to measure, but they can erode trust in democratic systems, as well as in public discourse and information integrity, and are often designed to inflame social divisions. They also have the potential to cause severe harm by inflaming tension and inciting violence in some settings. However, while it is essential that

platforms seek to counter the organized spread of disinformation, care must be taken to avoid limiting free expression through disproportionate and unnecessary action.

The Board selected these cases as a platform to talk about influence operations because the content and signals around it raised the possibility that each post could be part of one. After the cases were announced, Meta investigated and determined that neither post was part of an influence operation and the Board accepted that finding.

The remainder of this decision assesses account disablement under the Inauthentic Behavior and Account Integrity policies against Meta's human rights commitments.

III. Legality (Clarity and Accessibility of the Rules)

The principle of legality requires rules limiting expression to be accessible and clear, formulated with sufficient precision to enable an individual to regulate their conduct accordingly (General Comment No. 34, para. 25). Additionally, these rules "may not confer unfettered discretion for the restriction of freedom of expression on those charged with [their] execution" and must "provide sufficient guidance to those charged with their execution to enable them to ascertain what sorts of expression are properly restricted and what sorts are not" (*Ibid.*). The United Nations Special Rapporteur on freedom of expression has stated that when applied to private actors' governance of online speech, rules should be clear and specific (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 regarding their enforcement.

Meta uses AI systems and human expertise in its enforcement against influence operations, analyzing a wide variety of signals to identify and remove them. The Inauthentic Behavior policy establishes the public-facing rules in this area. However, it is cursory and lacking in detail. The policy states that inauthentic accounts and other assets cannot be created or used to deceive Meta or its users, evade enforcement of Meta's policies, or misuse Meta's systems to harass, intimidate or silence others, but does not provide specific examples of different types of prohibited behavior.

The Board notes that influence operations are carried out by highly sophisticated, well-resourced, and organized entities. Every detail given by Meta and other social media platforms about how they counter them will be analyzed and used by those entities to enhance their operations. The Board has previously challenged the use of this argument against transparency in different contexts (see, among others, Symbols Adopted by Dangerous Organizations). However, given the rapidly evolving nature of influence operations, including their growing use of AI, specific and detailed public-facing rules would require constant changes and would likely lead to abuse by the very entities platforms are trying to protect their users from. While international human rights law requires that the rules be clear and accessible to both users and those enforcing them, the principle does not require that platforms publicly share all details of the detection methods that underpin them, especially in situations where disclosure would undermine the efficacy of those mechanisms.

On that basis, the Board finds that the Inauthentic Behavior policy is sufficiently clear to meet international human rights norms. However, the policy could be improved by providing descriptions of the categories of behavior that the policy prohibits as well as an explanation that the enforcement is network-based rather than content-based.

The Account Integrity policy similarly does not provide sufficient examples or details of its network prohibitions. While the same caution against providing enforcement information to threat actors is relevant here, the Board highlights its previously raised concerns that Meta's public rules around account disablement are scattered and contradictory (see Account Ban for Targeting Public Figures).

IV. Legitimate Aim

Any restriction on freedom of expression should also pursue one or more of the legitimate aims listed in the ICCPR, which includes protecting the rights and reputations of others.

This is the first time the Board has assessed whether the Inauthentic Behavior policy serves a legitimate aim. Meta's policy rationale for the Inauthentic Behavior policy states that it targets behavior that has "the goal of deceiving Meta or our community or to evade enforcement under the Community Standards." Prohibiting this sort of activity pursues a legitimate aim by seeking to protect users' right to receive information while also protecting them from fraud. While the permissible aim of protecting "national security" stated in Article 19(3) does not directly apply to the aims of Meta because it is not a state, the Board recognizes that, as a private company, Meta has an analogous legitimate interest in preventing its platforms from being used by malicious actors to attack public institutions (Article 19(3) ICCPR).

The Account Integrity policy sets out the circumstances under which Meta will restrict or disable users' accounts when there are violations of the company's other policies and attempts to evade the penalties for those violations. Accordingly, any assessment of legitimate aim would need to focus on the underlying policy violations that Meta deemed warranted disablement. The penalty of account disablement could then be assessed under necessity and proportionality.

V. Necessity and Proportionality

Under ICCPR Article 19(3), necessity and proportionality requires that restrictions on 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 No. 34, para. 34).

Meta's prohibition of coordinated inauthentic behavior in the Inauthentic Behavior policy targets sophisticated networks of well-resourced and organized actors. In response to questions from the Board, Meta explained that there is a high evidentiary threshold for disabling a network for coordinated inauthentic behavior, with fewer than 300 networks being removed between 2017 and this case's announcement in 2026, according to Meta's threat reports. After staff or algorithms flag potential coordinated

inauthentic behavior, expert teams assess a broad range of technical signals, behavior across the suspected network of accounts and users, and the content posted, before determining whether the policy has been breached. Having reviewed the material shared by Meta, the majority of the Board is satisfied that, in theory, when this threshold is met as described, disablement of the accounts involved is likely to be necessary and proportionate, with no other less intrusive means available to protect the rights of others.

However, that does not necessarily mean that, in every specific instance, the policy will always be applied in a proportionate way. As the account in this case was not disabled under the Inauthentic Behavior policy, the Board cannot assess the necessity and proportionality of this specific disablement under this policy. Further, many of the behavioral signals that Meta relies upon to identify coordinated inauthentic behavior are not unique to the actors Meta targets with the policy. Newly created accounts, pseudonymous identities, geographic signals distorted by circumvention tools and tightly coordinated posting are also characteristic of activists, journalists and human rights defenders operating under state repression, for whom anonymity is a precondition of safety (A/HRC/29/32). Where enforcement systems are calibrated on adversarial network behavior, there is a serious risk of misclassifying organized protest as manufactured coordination, and it falls hardest on users with the least practical access to remedy.

When coordinated inauthentic behavior is found, lesser penalties and intermediate measures other than disablement of the account would not be adequate to address it. Suspensions and warnings will not dissuade organized threat actors who have invested significant resources into running an influence operation. While enforcement of this policy is, by necessity, slower and more time consuming than purely content-based moderation, intermediate steps are not appropriate. Automatically applying a warning label to content signaling to users that it may be part of an influence operation in advance of an investigation would give perpetrators time to repurpose elements of their networks. In addition, taking intermediate action against accounts before expert teams have been able to fully assess whether they meet the criteria for violating the

policy could result in a disproportionate number of false positive results, with serious consequences for mistakenly impacted users.

For a minority of the Board, while the approach Meta described having followed seems reasonable, there are practical concerns. The efficacy of the company's stated approach depends on the adequacy of the threshold and on the safeguards that are in place to ensure non-discriminatory enforcement, as well as on its resourcing for such investigations, with reported mounting pressures on various relevant streams that feed into this work, such as the Trusted Partners Program. Meta should therefore publicly disclose more information (or information should be provided to the Board and researchers) to independently assess the effectiveness of its current approach to the enforcement of its actor- or behavior-based policies, disaggregated by language, country and region.

The Board accepts Meta's finding that neither of the accounts in this case was part of a coordinated inauthentic behavior network, which would require links with other users acting in a coordinated fashion. However, the account that posted the first piece of content was found to be both inauthentic and connected to other accounts operated by the same user that had been disabled for violating Meta's policies. This was only uncovered by Meta after several steps – a user who reported the content appealed to Meta against its decision to keep it on the platform, then the user appealed to the Board, and the Board selected the case and asked Meta to explain why the account had been disabled. This suggests that, even though coordinated inauthentic behavior investigators are skilled specialists with sophisticated technology, there are potential gaps in how Meta discovers such content and sends it to those specialists for review. Some signals or patterns that humans can recognize may later be useful to authenticity-related investigations. The Board therefore proposes better communications channeling to surface potentially suspicious content.

The Inauthentic Behavior policy also prohibits the creation, use or claimed use of inauthentic accounts to misuse Meta reporting systems to harass, intimidate or silence others. The Board notes that coordinated networks of accounts are one of the main

delivery systems of harassment against a range of targets including activists, journalists, women and minority groups. While the Bullying and Harassment policy prohibits much of the concerning behavior, the Inauthentic Behavior policy prohibition is limited to misusing the reporting system and does not extend to other forms of coordinated harassment. The Board would welcome a case referral or policy advisory request that would allow it to explore how Meta prevents and responds to such coordinated harassment campaigns.

Regarding the Account Integrity policy, the necessity and proportionality of individual disablements will need to be assessed against the underlying policy violations that triggered each disablement. The relevant section of policy relied upon by Meta in this case targets accounts whose purpose is either solely or mainly to violate Meta's policies – when that condition is met, disablement is likely to be a necessary and proportionate response if enforcement is consistent, objective and non-discriminatory. Additionally, regarding accounts that have been created to replace those disabled for violating policies, those previous disablements were presumably necessary and proportionate themselves as well as being appealable. Therefore, this policy's rules against replacing such accounts function as an integrity measure to preserve those previous decisions. The Board does not approve or express any view on the underlying question of whether prior account removals were proper, but only the simple proposition that people should not be allowed to circumvent account removals by opening a new account.

VI. *Transparency*

Meta should consider regularly reporting the number of accounts taken down for spreading disinformation, including when those accounts are not part of a network. Meta's adversarial threat reports currently only address accounts taken down for networked behavior-based violations. There does not appear to be any recurring public reporting of account disablements for non-networked spreading of false or misleading content. It would be useful to researchers and civil society to understand how many accounts are taken down for spreading intentionally misleading content and this would not materially assist entities that engage in influence operations.

6. The Oversight Board's Decision

The Board upholds Meta's decision to leave up the first post, but reverses Meta's decision to leave up the second post.

7. Recommendations

Enforcement

1. To enhance information sharing and improve the investigation of influence operations, Meta should create a system for signal sharing between teams specializing in coordinated inauthentic behavior and at-scale moderators. This system should:

    - Be activated as part of any Crisis Policy Protocol designation involving countries where there is a particular risk of being impacted by foreign or domestic influence operations.
    - Enable coordinated inauthentic behavior review teams to flag content signals reviewers can use to escalate content that may be useful to their investigations. Content flagged under this system should not be removed unless it is found to be part of a coordinated inauthentic behavior campaign or violates another policy.
    - Allow at-scale reviewers to proactively report patterns that may warrant more comprehensive review when they encounter them on a case-by-case basis.

The Board will consider this recommendation implemented when Meta shares documentation detailing how this new system is triggered and enables these new information flows.

***Procedural Note:**

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

- Under its Charter, the Oversight Board may review appeals from users whose content Meta removed, appeals from users who reported content that Meta left up, and decisions that Meta refers to it (Charter Article 2, Section 1). The Board has binding authority to uphold or overturn Meta's content decisions (Charter Article 3, Section 5; Charter Article 4). The Board may issue non-binding recommendations that Meta is required to respond to (Charter Article 3, Section 4; Article 4). Where Meta commits to act on recommendations, the Board monitors their implementation.

- For this case decision, independent research was commissioned on behalf of the Board. The Board was 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.