Protest Footage Paired with Pro-Duterte Chants

2025-050-FB-UA

Summary

The Oversight Board has upheld Meta's decision to leave up a Facebook post that shared a manipulated video during a political crisis in the Philippines but notes that the content should have been labelled "High-Risk" because of the significant potential to deceive users on a matter of public importance. The Board recommends that Meta publicly explain its manipulated media labels and how it applies them. The video should have been prioritized for fact-checking, and the Board finds it is near-identical to other fact-checked content. Additionally, Meta should give fact-checkers better tools to address misleading content.

About the Case

In March 2025, former Philippine President Rodrigo Duterte was extradited to the International Criminal Court in the Netherlands to face charges for alleged crimes against humanity during his term in office from 2016 to 2022. A few days after the arrest, a Facebook user reshared a manipulated video that had been posted by another user. The reshared video contains footage from a Serbian protest that was unrelated to the Duterte arrest, with captions and audio added to make it appear to be a pro-Duterte demonstration taking place in the Netherlands.

The video had a text overlay that said, "Netherland." In the added audio, people are repeatedly chanting "Duterte", while the song, "Bayan Ko," plays in the Tagalog language. "Bayan Ko" was popular during anti-martial law protests in the Philippines in the 1980s.

The original post, which was shared hundreds of times and received about 100,000 views, was flagged by Meta's automated systems as possible misinformation. Meta included the content in the online queue for fact-checking. Separately, Meta temporarily lowered the visibility of the post in non-US users' Facebook feeds. Various similar videos went viral, and several were rated false by Meta's fact-checking partners in the Philippines. However, due to the high volume of posts in the queue, fact-checkers were not able to review this specific post. Another Facebook user reported the reshared post for spreading misinformation. Meta left the post up, after which the user appealed. A human reviewer considered the appeal and upheld the initial decision. The user then appealed to the Oversight Board.

Key Findings

The Board agrees with Meta that the post should have been left up because it did not include the types of content prohibited by Meta's Misinformation policy, such as discussing voting locations, processes or candidate eligibility. However, the Board notes that in addition to referring the content for fact-checking and temporarily showing the post lower on users' feeds, Meta should have applied a "High-Risk" label to the content because it contained a digitally altered, photorealistic video with a high risk of deceiving the public during a significant public event.

Given the importance of providing transparency around its manipulated media labelling, the Board recommends that Meta describe its different labels and criteria for applying them. Currently, the most detailed information on Meta's manipulated media labeling is in the Board's decisions.

Meta should have taken more steps to ensure the post was fact-checked. Although Meta prioritizes similar content for fact-checking during elections, the high-profile arrest of a former head of state, and other political crises that are "timely, trending and consequential," should be treated as critical events that qualify for heightened checks. Additionally, after its review, the Board finds the case content should also have

qualified as near-identical to previously fact-checked content and been labeled as such, yet recognizes that Meta faces challenges in making this determination at scale.

The Board notes that manipulated video can be part of concerted misinformation campaigns, in which similar, but not identical, content is posted and shared with subtle tweaks to evade fact-checking. This makes it imperative that Meta has robust processes to address viral misleading posts, including prioritizing identical or near-identical content for review, and applying all its relevant policies and related tools. Fact-checkers should also be given better tools to rapidly identify viral content that is likely to be repeating misleading claims.

The Oversight Board's Decision

The Board upholds Meta's decision to leave up the content.

The Board also recommends that Meta:

- Describe the different informative labels that the company uses for manipulated media, and when it applies them.
- Build a separate queue within the fact-checking interface that includes content that is similar, but not identical or near-identical, to content that has already been fact-checked in a given market.

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

Full Case Decision

1. Case Description and Background

After former Philippine President Rodrigo Duterte was arrested in March 2025 to face charges at the International Criminal Court (ICC) in the Netherlands, a Facebook user

reshared another user's video. The original video shows crowds gathered on a street holding signs and the Serbian flag. The video has text overlay stating, "Netherland," and a patriotic Tagalog song, "Bayan Ko," playing in the background. Audio of people repeatedly chanting "Duterte!" can also be heard. The song "Bayan Ko" was popular during the anti-martial law protests in the 1980s against the late dictator and former president, Ferdinand Marcos Sr. The original post had a caption in English: "Netherlands supporters." The reshared Facebook post, which is the subject of this case, has three pleading-face emojis as a caption. The post thus appears to be of a Serbian protest unrelated to Duterte's arrest, with captions and audio paired to it that make it appear to be a pro-Duterte demonstration in the Netherlands.

The original Facebook post has been shared hundreds of times and has received around 100,000 views since posting. When it was posted, a Meta classifier flagged the content as possible misinformation. The company then included the content in the online queue provided to its third-party fact-checking partners. Similar versions of this video went viral on Facebook in March 2025; several were fact-checked and rated as False by Meta's fact-checkers. Due to the sheer volume of such posts and their rapid spread on the platform, fact-checkers cannot rate all posts in this queue, which may include posts that are similar to content that they have already fact checked. According to Meta, fact-checkers did not rate the original post, and the "false" ratings applied to similar posts were not automatically applied here. Despite not being rated, Meta temporarily showed the case content lower in the Facebook feed of non-US users, including those in the Philippines, based on signals predicting it contained misinformation.

Days later, a user reported the reshared post, but according to Meta, the report was not prioritized for human review. Meta took no action on the report and closed it. After being notified that Meta left the content on the platform, the reporting user appealed Meta's decision. A human reviewer upheld the initial decision to keep the content up, which led the reporting user to appeal the decision to the Oversight Board. When the Board selected this case, Meta again added the post to a fact-checking queue, but it was not rated.

The Board notes the following context in reaching its decision.

During his term as president of the Philippines from 2016 to 2022, Rodrigo Duterte conducted a campaign against illegal drugs, characterized by excessive use of force and vigilante killings. The International Criminal Court (ICC) initiated a preliminary examination into the matter, followed by a full investigation. On March 11, 2025, Duterte was brought to the Netherlands to face charges for crimes against humanity at the ICC.

Current president Ferdinand Marcos, Jr., the son of the late dictator, won the 2022 national elections, with Sara Duterte, Rodrigo Duterte's daughter, winning the vice-presidency. Soon after winning the elections, the political alliance between them broke down. In 2024 and 2025, the political conflict escalated. Early this year, impeachment complaints were filed against Vice President Sara Duterte. In February, the Philippine House of Representatives voted to impeach Sara Duterte, which ultimately did not proceed. Duterte's arrest in March was seen by some as a result of their political feuding. Vice President Sara Duterte claimed the arrest amounted to a "kidnapping," while human rights activists and families of the slain drug war victims treated the arrest as a victory for accountability.

Shortly after Duterte's arrest, protests in support of Duterte erupted in the Philippines and the Netherlands. Activists and relatives of drug war victims also rallied, demanding accountability. During this time, multiple videos with footage of a Serbian rally presented as a pro-Duterte protest in the Netherlands, similar to the content under analysis, went viral on social media.

Three fact-checking organizations Meta works with in the Philippines rated some of these videos "False" based on signals that the protest footage depicted a protest in Serbia, not in the Netherlands, and was unrelated to former President Duterte's arrest. These fact-checkers also published their "False" ratings on their respective websites. According to AFP, these videos were "misusing unrelated visuals to inflate crowd sizes"

(PC-31360). Another similar video, which garnered over three million views on social media, showed a Catholic procession that was misrepresented as a pro-Duterte rally.

At the time it submitted a public comment for this case, AFP had fact-checked "more than 30 false or misleading claims" relating to Duterte's arrest (PC-31360). Rappler, another fact-checker for Meta in the Philippines, noted that "multiple versions of this specific manipulated video" had gone viral on Facebook as part of a "resurgence of information operations" and was part of an "inauthentic and organized campaign" intended to "manipulate public opinion and behavior in his favor" (PC-31349). Experts consulted by the Board stated that since 2016, fact-checkers and civil society groups have borne the brunt of combating disinformation.

More broadly, the use of social media in the Philippines is one of the highest in the world, with Facebook being the most used platform. Social media plays a significant role in shaping public discourse surrounding current events. For instance, in the 2016 national elections, advertising and public relations strategists functioned as "architects of networked disinformation" to manage political campaigns, hiring online influencers and trolls to seed specific political narratives. In the 2022 national elections, historical revisionism about the Marcos family's legacy spread on social media, which some believe contributed to Marcos Jr.'s presidential victory. Experts consulted by the Board explained that "social media influencers, vloggers, and digital workers from advertising and public relations firms have played significant roles in the production and dissemination of disinformation on social media platforms." As the rivalry between President Marcos and Vice President Sara Duterte intensified and gave rise to competing political narratives online, Filipinos became even more concerned about disinformation, with a study revealing a "record 67%" of people had increased concern over disinformation by mid-2025. Earlier this year, the Philippine Congress initiated an inquiry to address disinformation on social media, inviting influencers, vloggers and representatives of social media companies to testify.

Both under the prior Duterte government and current Marcos government, there have been serious challenges to freedom of expression in the Philippines. Journalists and

media outlets deemed critical of Duterte were attacked or shut down by the government during his presidency and targeted by online trolls. "Red-tagging," or the act of branding groups or individuals as "supporters, recruiters or members of the New People's Army or the Communist Party of the Philippines" became common practice, often targeting journalists and human rights defenders. Threats against the media continue under President Marcos Jr. In June 2025, the UN Special Rapporteur on freedom of expression called on the Philippine government to address the intimidation and harassment of journalists and human rights defenders in the country. These challenges, along with the rise of disinformation campaigns on social media, have eroded public trust in media institutions in the Philippines.

2. User Submissions

According to the reporting user, the case content purports to show a rally in support of Duterte following his ICC detention, but in fact it did not take place. The user stated that the content was of a different event and was therefore "misleading." They further state that the content was "already verified" by one of Meta's fact-checkers in the Philippines.

3. Meta's Content Policies and Submissions

I. Meta's Content Policies

a. Misinformation Community Standard

Content Subject to Removal

Under its Misinformation Community Standard, Meta removes misinformation where it is likely to directly contribute to the risk of imminent physical harm or to "interference with the functioning of political processes [voter or census interference]." Misinformation involving voter interference subject to removal is defined by a list of

prohibited misinformation about voting schedules or location, and voter or candidate eligibility.

Content Eligible for Third-Party Fact-Checking

For other types of political misinformation, Meta states it focuses not on removal but on "reducing its prevalence or creating an environment that fosters a productive dialogue." In the United States, Meta has discontinued its third-party fact-checking program and now addresses misleading content through Community Notes. Outside the United States, fact-checking remains available. For this purpose, Meta partners with third-party fact checking organizations certified through the non-partisan International Fact-Checking Network (or the European Fact-Checking Standards Network in Europe) to review and rate the accuracy of the most viral content on Meta's platforms.

Fact-checkers can review and rate the accuracy of public Facebook, Instagram, and Threads posts, including ads, articles, photos, videos, Reels, audio and text-only posts. Rating options are False, Altered, Partly False, Missing Context, Satire and True.

Fact-checkers decide what content to review, and complete their fact check independently from Meta. They may either identify content on their own initiative or select from a queue of Meta referrals of potential misinformation. Meta refers content based on various signals, including "how people are responding," whether users flag a piece of content as "false information," or when comments on a post "express disbelief" about its authenticity.

Fact-checkers prioritize viral false information and verifiably false claims that are "timely, trending and consequential" in their relevant country and language. Content ineligible for fact-checking includes "content that doesn't include a verifiable claim," "opinion and speech from politicians," and "digitally created or edited media containing one of Meta's AI transparency labels or watermarks on the basis of its authenticity." However, when manipulated media contains a false claim separate from the use of digitally created or edited media, fact-checkers may still rate the post. Fact-

checked content rated False, Altered or Partly False may be demoted, not recommended and rejected for ads.

Content Eligible for Manipulated Media Labeling

Under the manipulated media rules of the Misinformation policy, misleading content that is "digitally created or altered" but "does not otherwise violate other Community Standards" may receive an informative label "on the face of the content" or be rejected if submitted as an advertisement. This applies to a "photorealistic image or video, or realistic sounding audio, that was digitally created or altered and creates a particularly high risk of materially deceiving the public on a matter of public importance." Meta does not include in the policy further criteria for when it "may" apply this label.

Meta requires people to disclose, through its "AI-disclosure tool," whenever users post "organic content with photorealistic video or realistic-sounding audio that was digitally created or altered." Meta states it may apply penalties if users fail to do so.

Meta applies three different informative labels for manipulated media: the "High-Risk" label, the "AI Info" label and the "High-Risk AI" label. In the Alleged Audio Call to Rig Elections in Iraqi Kurdistan case, Meta informed the Board that to apply a High-Risk label, the content must (i) create a particularly high risk of materially deceiving the public on a matter of public importance; and (ii) there are reliable indicators that the content was digitally created or altered. The High-Risk AI label is similar to the High-Risk label, but for content that has reliable indicators of being created or altered with AI. The "AI Info" label is for images, not video or audio, made with AI that Meta detects through "industry standard AI image indicators or when people disclosed that they were uploading AI-generated content."

Applying a High-Risk label does not result in the demotion of the content or its removal from recommendations. Instead, when users attempt to reshare a High-Risk-labeled post on Facebook, Instagram or Threads, they receive a pop-up notice alerting them that the content may have been digitally created or altered.

b. Demotion policies

Meta's Content Distribution Guidelines describe the types of content that may be demoted for being "problematic" or "low quality." Meta bases its demotion of these types of content on its "commitment to the values of Responding to People's Direct Feedback, Incentivizing Publishers to Invest in High-Quality Content and Fostering a Safer Community." According to Meta, this enables people to share content "without being disrupted by problematic or low-quality content."

Among the types of content that Meta states it demotes are fact-checked misinformation as well as content assessed as "likely violating" Meta's Community Standards. When content is posted on Facebook, it is assessed by an array of classifiers to determine whether it violates one of the platform's Community Standards. These classifiers report a degree of "confidence" in the likelihood of a violation that the company uses as a signal to take certain actions. When that confidence score for certain content is high, but not sufficiently high to trigger immediate removal, Meta may demote the content. If the content is found to violate any of the covered policies, it is removed. The company may adjust the confidence threshold that individual classifiers require to take certain actions, such as in times of crisis.

II. Meta's Submissions

Meta states it is not possible for fact-checkers to review all potential misinformation, so the company instructs fact-checkers to prioritize "viral and relevant topics that have the potential to cause harm or spread quickly."

Meta told the Board it relies on two methods to prioritize content for fact-checking during critical times. First, Meta can mark certain enqueued content as urgent. Second, during high-priority global events such as certain crises or elections that Meta considers "Trending Events," Meta filters and enqueues related content for fact-checking. This filter uses a list of relevant keywords identified by local market operations specialists

with language and contextual expertise using defined guidelines. Content identified through this process receives a topic label with the corresponding event name and is included in a dedicated, time-bound Trending Events filter view within the fact-checking tool. To qualify as a Trending Event, an issue must be important (such as a real-world crisis, civic debate or risk of offline harm) and considered at high risk for misinformation spread. Fact-checkers retain discretion as to whether to select urgent or trending content for review.

In addition to selecting content from Meta's queue, fact-checkers can identify content for review independently on Meta's platforms. Meta told the Board that fact-checkers have access to the Meta Content Library, a web-based tool that allows those who have access to perform systematic searches of some publicly accessible content on Meta's platforms. Fact-checkers can select content discovered in the Meta Content Library for fact-checking. Meta stated it "proactively invites" fact-checkers to use this tool.

Meta did not treat the ICC arrest of the former president as a trending event and did not certify the content as urgent for fact-checking. The company told the Board it did not receive information about misinformation trends leading to imminent violence or physical harm in connection with the May 2025 Philippine midterm elections. However, the company also shared that its fact-checking partners surfaced "prevalent misinformation claims such as false claims relating to former President Rodrigo Duterte's arrest by the ICC and his daughter [Vice President] Sara Duterte's impeachment."

Meta explained that once a fact-checker rates a piece of content, they tell Meta where the false information is in the content (whether in the caption, only in the audio or video or part of it, or when both the media and caption are considered together). Meta then uses matching technology to apply that rating only to content that is identical or near-identical for that specific aspect (i.e., the same media, or the same media paired with the same caption). If the fact-checker rates only the video in the post, Meta will apply the rating to any post that contains the same or nearly the same video. However, if the

fact-checker rates both the video and caption considered together, the rating will only be applied to posts with both the same or nearly the same video and caption.

Meta defines "identical" content to describe content sharing as the "exact same attributes" as media rated by fact-checkers. "Near-identical" and "almost exactly the same" content has "minor variations in formatting or overlay text, but conveys the same debunked claim." Meta does not automatically apply labels at scale if content does not meet the standards for identical or near-identical content. This is because the company finds that small differences in how a claim is phrased can affect whether it is true or false. For Meta, this approach "helps prevent incorrect application of fact-check ratings."

In the present case, Meta found that the case content did not meet the identical or near-identical threshold to apply the same rating label that Meta's fact-checkers had applied for similar videos portraying the protest in Serbia as a pro-Duterte protest in the Netherlands. For instance, some videos that were rated by fact-checkers had different audio, as this post had the song "Bayan Ko" in the background, or the captions were different.

Content referred to fact-checkers for review times out of the referral queue after seven days if fact-checkers do not select the content for review. Meta explained that this time frame reflects the trend that "most views happen within the first few days of a content's lifecycle." Fact-checkers may still select content to review outside of Meta's referral queue, and rate content "regardless how long it has been on the platform. For example, while fact-checkers generally focus on timely topics, older content may resurface in light of new issues."

With respect to the manipulated media rules of its Misinformation policy, Meta did not apply any of its labels, including the High-Risk label, to the case content. Meta told the Board this was because it was not escalated internally. When it was analyzed following the Board's selection of the case, the content was "already more than 2 months old." Given "the content's age and lack of virality," Meta decided not to apply the

manipulated media label at that point either. Meta emphasized that the High-Risk label is an escalation-only policy. It is used sparingly to "address digitally created content that poses an especially acute risk of misleading the public about an important issue at a critical time." According to Meta, this includes "when content is posted close in time to a critical event, such as an election, and there is not enough time for the information ecosystem or fact-checkers to address the content at issue." Because the content was posted two months before the May 2025 midterm elections in the Philippines, for Meta this was sufficient time for the "information ecosystem or counter-speech to correct any misinformation in the post (and here, similar versions of the video were fact-checked)." It appears that Meta did not consider either Duterte's arrest or the subsequent protests to independently be "critical events."

The only enforcement action that Meta took on this content was to temporarily lower the post's visibility in the Facebook feed of non-US users, including in the Philippines, based on signals predicting that the post contained misinformation.

The Board asked questions on how fact-checking works, how Meta allocates resources to fact-checkers, how similar videos on Duterte's arrest were compared to the reshared video, and whether Meta considered the manipulated media rules of the Misinformation policy in moderating the case content. Meta responded to all questions.

4. Public Comments

The Oversight Board received six public comments that met the terms for submission. Five of the comments were submitted from Asia-Pacific and Oceania and one from Europe. To read public comments submitted with consent to publish, click here.

The submissions covered the following themes: the constraints faced by fact-checkers, fact-checkers' local expertise, the inadequacies of Meta's current approach to addressing coordinated disinformation campaigns, and the accounts responsible for sharing misinformation in the Philippines.

In July 2025, as part of ongoing stakeholder engagement, the Board consulted with representatives of fact-checking organizations, academics and other misinformation experts. The roundtable discussed how fact-checking works in practice, Meta's fact-checking outside the United States, the volume of potentially misleading content that fact-checkers face and the ability of those spreading disinformation to make changes to the post to evade detection, as well as the risk of overenforcement posed by matching technology to detect identical and near-identical content.

5. Oversight Board Analysis

Disinformation campaigns pose threats to information integrity, public trust and democracy itself. The Board selected this case to examine how Meta addresses false or misleading information on its platforms, especially when shared during moments of heightened political tension and in contexts where disinformation influences public debate. The Board analyzed Meta's decision in this case against Meta's content policies, values and human rights responsibilities. The Board also assessed the implications of this case for Meta's broader approach to content governance.

5.1 Compliance With Meta's Content Policies

I. Content Rules

The Board finds that the case content does not meet the criteria for removal under the Misinformation policy standard. It does not provide information about any of the prohibited categories under this policy line, including voting or census locations, voting processes, or voter or candidate eligibility. The Board also finds that the case content meets the eligibility criteria for fact-checking under Meta's Misinformation policy, and agrees with Meta for submitting the post for fact-checking. The case content meets the eligibility criteria for manipulated media labeling under Meta's Misinformation policy. Therefore, Meta should have applied a "High-Risk" label to the content.

A. Misinformation Community Standard and Third-Party Fact-Checking

The case content meets the eligibility criteria for fact-checking under Meta's Misinformation policy. The post contains a verifiable claim, does not involve speech from politicians, and does not have an AI transparency label that would make the content ineligible for fact-checking (i.e., the misleading nature of the post is in how it was digitally created or altered, not what the post claims). It also satisfies one of the elements for content to be prioritized for fact-checking — it is "timely, trending and consequential." The post was shared shortly after former President Duterte had been arrested, when the circumstances of his arrest were being contested online and offline. At the time the case content was posted, protests had broken out both in support of and against Duterte's arrest. Posts similar to the case content claiming that pro-Duterte protests were taking place were going viral on social media at the time. One of Meta's fact-checkers reported that similar videos garnered more than three million views. Moreover, the arrest took place against a broader context of intense political feuding between President Marcos and Vice President Sara Duterte, in a country where disinformation influences political debate.

Meta should have taken action to prioritize such content in the queue it provides fact checkers. The company should have marked the case content as urgent when it enqueued the content for fact-checking. Meta should also have designated potential misinformation surrounding former President Duterte's arrest as a Trending Event and proactively identified and surfaced similar content after Meta's fact-checkers rated similar videos false. In line with Meta's own criteria, the issue is important for "civic debate" and there was a "high risk for misinformation spread." It appears from the information provided to the Board that Meta was remiss in this regard.

The case content should also have qualified as near-identical to previously fact-checked content and been labelled as such. Fact-checkers who had rated similar videos identified the inaccuracy in the pairing of the protest footage in Serbia with audio of people chanting "Duterte!" in the background. Those rated videos did not contain the song "Bayan Ko" in addition to the chanting. Because the case content here included

this song, it was sufficiently different for Meta's matching technology not to identify it as identical or near-identical to similar videos previously rated False.

For the Board, the changes introduced constitute "minor variations," and the content repeats "the same debunked claim" as similar videos rated false by Meta fact-checkers. The minor differences in audio, text and caption did not change the fundamental misleading characteristics of this content. When subtle changes do not affect the misleading nature, it should be deemed near-identical.

The Board also recognizes that the automated systems Meta currently uses may be incapable of making this determination at scale. Applying the label to content with different audio without any evaluation could potentially lead to labeling content with audio criticizing or debunking the misleading video. Stakeholders consulted by the Board noted that matching technology would not be able to anticipate all possible uses and nuances of expression, and that expanding it, without developing other tools to address misinformation, could lead to overenforcement. At the same time, Meta's current approach is not effective in responding to actors who tweak content in subtle ways to game the system and evade enforcement. Meta should allocate sufficient resources, automated or otherwise, to ensure that fact-checkers' ratings are applied effectively.

B. Misinformation Community Standard and Manipulated Media

The Board finds that the case content amounted to manipulated media that should have received a label. Based on similar public fact checks, the content is a "photorealistic video" that appears to be "digitally created or altered." The Board considers the video footage of the Serbian protest as meeting the "photorealistic" requirement of the Manipulated Media policy. The pairing of the protest footage in Serbia with audio of people chanting Duterte's name and adding a text overlay "Netherland" to it is a digital alteration of the original footage. The text overlay "Netherland" further distorts the meaning of the video by making it appear that the protest took place in the Netherlands, when in fact, it happened in Serbia. The post also

“does not otherwise violate other Community Standards” that would merit a different treatment, such as removal. The content poses a "particularly high risk of materially deceiving the public on a matter of public importance” as it concerns a significant political issue in Philippine politics. Similar videos were circulating on Meta’s platforms and had been rated False by Meta’s fact-checkers in the Philippines. The volume of similar misleading content thus increased the risk of materially deceiving the public. Based on these factors, Meta should have applied its High-Risk label.

Meta argued it did not apply any manipulated media label as the content was not escalated to Meta’s internal teams, and it was posted far enough in advance of the May 2025 elections such that sufficient counter-speech debunking the false claim could occur. Moreover, at the time the Board selected the case, Meta did not consider adding the label given the post’s age and the lack of virality.

The Board disagrees with Meta. First, these justifications do not seem to have any basis in Meta’s public-facing policies. Second, even using these internal criteria, the Board reaches different conclusions.

Meta should have treated the arrest of former President Duterte as a critical event that warranted escalation of relevant posts for possible application of the High-Risk label. Under Meta’s own standard, the time of Duterte’s arrest represented a tense political moment in the Philippines, generating critical public discourse about the circumstances of the arrest and protests in support of and against the former president in the Philippines and abroad. Stakeholders consulted by the Board highlighted the rapid nature with which narratives concerning Duterte’s arrest proliferated on Meta’s platforms. As Meta noted, similar versions of the video were fact-checked. The arrest also happened during longstanding political feuding between the two highest government officials of the country, with Vice President Sara Duterte alleging her father was kidnapped.

Based on its actions here, Meta appears to limit its interpretation of critical events eligible for manipulated media labeling to elections, excluding other political crises.

The Board considers the arrest and subsequent extradition to the ICC of a former head of state to be a critical event. Given the political situation in the Philippines at the time of the former president's arrest, it was a mistake for Meta to treat the upcoming May 2025 elections as the only critical event at play. Therefore, by effectively limiting the designation of a critical event to elections, Meta undermined the utility of its own rules to help limit the spread of manipulated media during situations of heightened political tension. The content that the manipulated media rules cover in practice seems significantly narrower than what the policy appears to convey.

Even assuming that the critical event in this case was the May 2025 elections, Meta should not have solely relied on its fact-checking partners' ability to surface and rate misinformation claims. Fact-checkers consulted by the Board as well as public comment submissions overwhelmingly highlighted the severe resource constraints that fact-checkers face in doing their work (PC-31362, PressOne (Philippines); (PC-31357, Foundation for Media Alternatives (Philippines)). Other fact-checkers explained how they are unable to rate all possible versions of similar videos. Along with the volume of content, purveyors of disinformation make subtle tweaks to content to distinguish new posts from previously fact-checked content and accordingly evade Meta's matching technology for identical and near-identical content (PC-31357, Foundation for Media Alternatives (Philippines); PC-31349, Rappler (Philippines)). Stakeholders emphasized that the continued proliferation of misleading posts without labels on Meta's platforms tends to drown out factual claims (PC-31358, European Fact-Checking Standards Network). One of Meta's fact-checkers noted that these posts form part of a broader organized campaign to spread disinformation online (PC-31349, Rappler (Philippines)). Given these circumstances, Meta should have concluded that "there [was] not enough time for the information ecosystem or fact-checkers to address the content at issue" which would have led to prioritizing this kind of content for fact-checking.

When misleading information seems to form a part of a broader systemic disinformation campaign to influence public opinion about a particular political or social issue under heightened tensions, it is even more necessary to address viral

misleading posts, especially when proliferation strategies quickly evolve to evade detection and review. In cases like this, Meta should apply all its relevant policies (including fact-checking and labeling) and related tools.

II. Enforcement Action

In the Altered Video of President Biden decision, the Board expressed its concern about Meta's practice of demoting content that third-party fact-checkers rate as "false" or "altered," without informing users or providing appeal mechanisms. The Board stated in that decision that "[d]emoting content has significant negative impacts on freedom of expression. Meta should examine these policies to ensure that they clearly define why and when content is demoted, and provide users with access to an effective remedy (Article 2 of the ICCPR)." At the same time, the Board recognizes that there may be instances where demotion may be warranted as a less intrusive measure than content removal (see e.g., Posts Supporting UK Riots, Criminal Allegations Based on Nationality, Iranian Make-up Video for a Child Marriage).

In the present case, Meta's action to temporarily show the case content lower in feed for non-US users, including those in the Philippines, aligned with its practice of demoting borderline policy-violating content. As the Board also finds the content to be nearly identical to content previously rated false, it would also be demoted under Meta's policy to demote content that third-party fact-checkers rate false. However, the Board also reiterates its deep concern regarding the lack of clarity in demotion policies, appeal opportunities, and the potential impact on political expression.

5.2 Compliance With Meta's Human Rights Responsibilities

The Board finds that keeping the content on the platform, with a High-Risk manipulated media label, as Meta's own policies require, would have been consistent with Meta's human rights responsibilities. The Board agrees that Meta's referral of the content for fact-checking aligned with its human rights responsibilities, and in this case, improving

the tools available to third-party fact checkers to enable their review would be one way to fulfill those responsibilities.

Freedom of Expression (Article 19 ICCPR)

Article 19 of the ICCPR provides for broad protection of expression, including political discourse (General Comment No. 34, paras. 11-12). It provides "particularly high" protection for "public debate concerning public figures in the political domain and public institutions" (General Comment No. 34, para. 38 20; see also General Comment No. 25, para. 12 and 25). The UN Human Rights Committee has emphasized that freedom of expression is essential for the conduct of public affairs and the effective exercise of the right to vote (General Comment No. 34, para. 20; see also General Comment No. 25, paras. 12 and 25).

Mere falsehood cannot be the sole basis of removing speech under international human rights law (UN Report of the Secretary-General on countering disinformation, A/77/287, para. 13). It can only be restricted if it passes the three-part test of legality, legitimacy, and necessity and proportionality. In 2024, several UN Special Rapporteurs and UN Working Groups made a joint statement on strengthening democracy and human rights during worldwide elections: "Social media companies should review and make transparent their processes of content moderation and algorithms to ensure they do not contribute to censoring dissent and promoting misinformation. Tech companies should carry out due diligence, invest in fact-checking, and understanding of local languages and local contexts in their content moderation policies."

When restrictions on expression are imposed by a state, they must meet the requirements of legality, legitimate aim, and necessity and proportionality (Article 19, para. 3, ICCPR). These requirements are often referred to as the "three-part test." The Board uses this framework to interpret Meta's human rights responsibilities in line with the UN Guiding Principles on Business and Human Rights, which Meta itself has committed to in its Corporate Human Rights Policy. The Board does this both in relation to the individual content decision under review and what this says about Meta's

broader approach to content governance. As the UN Special Rapporteur on freedom of expression has stated, although "companies do not have the obligations of Governments, their impact is of a sort that requires them to assess the same kind of questions about protecting their users' right to freedom of expression" (A/74/486, para. 41).

I. 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 UN 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's Misinformation Community Standard and related policies are clear as applied to the content in this case, but should be improved. The public-facing language of the Misinformation policy clearly apprises users of the applicable rules and the different consequences of posting misinformation and manipulated media on Meta's platforms (i.e., content is removed, fact-checked, or a manipulated media label applied to it). Meanwhile, the types of content eligible for fact-checking, criteria for appointing fact-checkers, and demotion policies can be found on different pages of the Transparency Center.

In the Alleged Audio Call to Rig Elections in Iraqi Kurdistan decision, the Board stated that Meta should consider "integrating the information on all the different manipulated media labels on one page in the Transparency Center so that users can easily learn more about them," and reiterates this guidance here.

Moreover, the Board is concerned that Meta's interpretation of its Manipulated Media rules for AI labeling results in a narrower scope than the public policy implies. Meta should describe the different types of AI labels it applies, the criteria to apply them and their consequences. Currently, the most detailed description of the three AI labels Meta uses is in the Board's recent Alleged Audio Call to Rig Elections in Iraqi Kurdistan decision. The public-facing rules should reflect Meta's internal rule that a critical event is required to apply a High-Risk label to a piece of content, as well as what qualifies as a critical event.

II. 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 of others (Article 19, para. 3, ICCPR). In previous decisions, the Board held that protecting the right to participate in public affairs (Article 25, ICCPR) is a legitimate aim for Meta's Misinformation policy (Alleged Audio Call to Rig Elections in Iraqi Kurdistan, Altered Video of President Biden).

III. Necessity and Proportionality

Under ICCPR Article 19(3), necessity and proportionality require 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). For this purpose, the company can first evaluate the tools it has to achieve the legitimate aim without burdening speech; second, if this is not possible, identify the tool that least intrudes on speech; and third, assess whether the measure it selects is effective (A/74/486, para. 52).

In assessing the necessity and proportionality of these measures, as well as in determining what other measures are necessary, the Board considered the following:

a) that the content was posted during heightened political tension in the Philippines, with varying narratives circulating about the circumstances of former President Duterte's arrest; b) high social media usage in the country; c) a history of disinformation and misinformation in the country, particularly in the lead-up to and during elections; d) the generally polarized nature of the political and media environment; e) the continued delegitimizing of news media and journalists as well as the decline in public trust in media institutions; f) the likelihood that the media contained in the post is altered, as indicated by similar posts rated False by all three of Meta's fact-checkers in the country; and g) the likelihood that digitally altered media will mislead and influence the public on an issue that has fueled political polarization in the country.

The Board notes Meta's actions with respect to the case content: submitting the case content for fact-checking, first, shortly after the content was posted, and again when the Board selected this case; and temporarily showing the content lower in the Facebook feed of non-US users, including in the Philippines, based on signals predicting that the content contained misinformation.

While Meta took some actions to facilitate fact-checkers reviewing the content, it should have taken further steps in accordance with its policies. For example, the company should have tagged the case content as urgent when it enqueued it for fact-checking, and treated the arrest as a Trending Event involving civic debate. Meta also should have proactively identified and surfaced content like this, given that Meta's fact-checkers flagged prevalent misinformation claims about former president Duterte's arrest, and similar content had been rated False by Meta's fact-checkers.

Given the nature of political disinformation in contexts such as the Philippines described above and its rapid spread, Meta should explore ways to better identify and address misinformation through all available approaches. As the UN Special Rapporteurs' joint statement emphasized, companies should respond to misinformation with content moderation practices informed by local context.

Meta should improve its mechanisms for fact-checkers to review similar content to previously fact-checked content. In the Altered Video of President Biden decision, the Board noted that applying a label to a small portion of content "could create the false impression that non-labeled content is inherently trustworthy." Currently, fact-checkers may independently identify content for fact-checking, in theory allowing content similar to already-labeled misleading information to be addressed. However, stakeholders have expressed to the Board that CrowdTangle, the Meta-owned transparency tool deprecated in 2024, was integral to this process. Although the Meta Content Library is available to fact-checkers, the actual utility of this tool to fact-checkers is uncertain (See Posts That Include "From the River to the Sea"). Stakeholders noted that the ability of journalists to access the tool is inconsistent, while some noted that the tool is not easy to use for fact-checking work, citing an inability to search for text in videos, reverse-image search and search for public groups by location. Based on this input, Meta should continue to proactively engage with third-party fact-checkers to ensure the Meta Content Library is well-suited for their purposes.

The Meta Content Library must also be complemented by other measures to address prevalent misinformation claims. For this purpose, the Board recommends that Meta develop tooling in its fact-checking queue that allows fact checking partners to rapidly identify viral content that has not qualified as identical or near-identical, yet likely repeats rated false or misleading claims. This will surface misinformation claims similar to those fact-checked while avoiding the risk of overenforcement from broadening the definition of "identical" and "near-identical" content.

Meta should take steps to improve its misinformation response, in part to avoid placing the burden on addressing misleading information on its third-party fact checking partners. Despite Meta's statement that it is making investments in technology to better detect "subtle distinctions in content" that may share misleading information, the current case shows the need for further improvement. For its existing fact-checking program, for example, Meta should ensure partners are supported and resourced to perform the challenging work that Meta counts on them to provide. Meta should also

address accounts that spread misinformation repeatedly, including by enforcing its Inauthentic Behavior policy line on coordinated inauthentic behavior.

The Board is also concerned with Meta's failure to apply the High-Risk label to the case content, as the company's policies allow. The Board finds that this would likely have helped prevent the further spread of false or misleading information in the Philippines during that critical time. Users seeing the post would have been alerted that the content may have been digitally altered, due to the audio of people repeatedly chanting "Duterte!" and the song "Bayan Ko" in the background being paired with the video footage of the protest in Serbia. Notably, the consequence of a High-Risk label is to show a pop-up to users who want to reshare the post, that the post they intend to share may be digitally altered. While the label is informative, this notice creates a degree of friction that can help reduce the spread of misleading posts about current events when this type of content is likely to peak. Similarly, in the Alleged Audio Call to Rig Elections in Iraqi Kurdistan decision, the Board took issue with Meta's selective application of its High-Risk label prior to a polarized election in Iraqi Kurdistan, further highlighting the confusion that the uneven labeling of misleading content can cause. The Board notes reporting observing inconsistency in the application of AI labels to content, even when industry-standard signals of AI generation are present. As AI-generated content becomes more popular on social media, it is critical that Meta improve its policy enforcement in this area.

In this case, the Board is seriously concerned that Meta did not apply the High-Risk label. Meta should have done that here, complementing its fact-checking program by leveraging other policies at its disposal.

6. The Oversight Board's Decision

The Oversight Board upholds Meta's decision to leave up the content.

7. Recommendations

Content Policy

1. To better inform users of how the Misinformation Community Standard manipulated media policy is enforced, Meta should explain the different informative labels that Meta uses for manipulated media, including that the High-Risk label is applied in relation to a critical event, and what counts as a critical event.

   The Board will consider this recommendation implemented when Meta updates the language in the Misinformation Community Standard to reflect the change.

Enforcement

2. To enable third-party fact-checkers to efficiently address patterns of misinformation, Meta should build a separate queue within the fact-checking interface that includes content similar, but not identical or near-identical to content already fact-checked in a given market.

   The Board will consider this recommendation implemented when Meta shares information with the Board detailing this new interface feature and how it enables fact checkers to incorporate new, similar content into existing fact checks.

***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.*