AI-Generated Video in Israel-Iran Conflict

2026-004 FB-UA

Summary

In analyzing the spread of AI-generated content in armed conflicts in a case on the 2025 Israel-Iran war, the Oversight Board calls on Meta to do more to allow users to identify such output. Its approach to surfacing AI-generated content must evolve. This includes providing details at scale about the origin of media, based on content provenance standards, investing in stronger detection tools and developing better methods for appropriate labeling. Meta needs to create a new, separate set of rules to ensure users can reliably recognize AI-generated content. Additionally, it should amend its current policies to ensure a timely and adequate response to deceptive AI-generated output.

The company needs to meet its public commitments and employ its own tools and others available across the industry to effectively address deceptive generative AI content that spreads among platforms.

The Board overturns Meta's decision to leave up the post in this case without a High Risk AI label.

Why This Matters

As the quantity and quality of AI-generated content increase, its impact on people and societies will be profound. The risks are heightened when deepfake output designed to deceive, manipulate or increase engagement is shared during conflicts and crises, such as in Iran and Venezuela in 2026, and spreads rapidly on different companies' platforms. During those two crises, there were claims that deceptive AI-generated content was authentic and that authentic content was fabricated. That heightens the public's inability to discern truth, emblematic of the liar's dividend, leading to a general distrust of all information. AI-driven influence campaigns are a growing challenge seen globally in recent years, exacerbated in restrictive media and internet ecosystems that limit credible information. However, AI-generated output being misleading is not in itself a

legitimate reason to restrict freedom of expression. The industry needs coherence in helping users distinguish deceptive AI-generated content and platforms should address abusive accounts and pages sharing such output.

About the Cases

The Israel-Iran war in June 2025 signaled an inflection point, with the presence of deceptive generative AI content on social media becoming known as its own "soft war." Such deceptive output was reported as garnering huge numbers of views, and both Israeli and Iranian governments were accused of AI-driven influence attempts. On June 15, two days into the 12-day Israel-Iran conflict, a video was posted to a Facebook page that claimed to be a news source. The posting user was in the Philippines. The video depicted extensive damage to buildings with overlaid text in English reading "Live now – Haifa Towards Down" [sic] with the posting date. The video was very similar to one originating on TikTok and identified by an independent fact-checker (Agence France-Presse) as false and AI-generated. A caption on the Facebook post listed many headline-style phrases linked to the conflict and unrelated terms and hashtags. The post received more than 700,000 views, with several comments noting that the content was AI-generated.

Six users reported the case to Meta, but it was neither reviewed by the company nor checked by third-party fact-checkers. A user appealed to the Board. After the Board selected this case, Meta confirmed the post did not violate the Misinformation Community Standard because it did not "directly contribute to the risk of imminent physical harm," and did not require an AI label.

Obvious signals of deception related to the post led the Board to question Meta over the identity and behavior of accounts linked to the page. The company subsequently disabled three accounts linked to the page for engagement abuse and inauthenticity, removing the page and, with it, the case content. The page had been eligible to monetize through Meta's Stars program.

Key Findings

The Board finds that the content posed a material risk of misleading the public on an important matter at a critical time, so Meta should have applied a “High Risk AI” label. The post did not meet the threshold for removal (posing a risk of imminent physical harm or violence). Meta must do more to address the proliferation of deceptive AI-generated content on its platforms, including by inauthentic or abusive networks of accounts and pages, particularly on matters of public interest, so that users can distinguish between what is real and fake.

The Board is concerned by reports that Meta is inconsistently implementing Coalition for Content Provenance and Authenticity (C2PA) standards even on content generated by its own AI tools, and that only a portion of such output receives proper labeling. The C2PA sets out technical standards to embed provenance information as metadata in content, allowing platforms to more easily identify AI-generated content and apply labels to inform users.

The current mechanisms for affixing even the standard label of AI Info to video (user self-disclosure or an escalation to the Content Policy team) are neither robust nor comprehensive enough to contend with the scale and velocity of AI-generated content, particularly during a crisis or conflict where there is heightened engagement on the platform. A system overly dependent on self-disclosure of AI usage and escalated review (which occurs infrequently) to properly label this output cannot meet the challenges posed in the current environment. Some Board Members additionally noted that High Risk AI labels (for output that could deceive people on important matters) must also be coupled with demotion or removal from recommendations to address concerns over spreading the impact of deceptive content.

Meta’s narrow approach to fanning out ratings to identical and near-identical content may have meant this post did not receive a fact-checking rating. Resource constraints and a remarkable volume of output make it difficult for fact-checkers to ensure timely review of all deceptive content, especially during a conflict or crisis. The Board reiterates that Meta should ensure that fact-checkers are adequately resourced and have guidance on prioritizing content from conflicts. The Crisis Policy Protocol (CPP)

and Trending Events designations should have allowed Meta to ensure more effective support for third-party fact-checkers during the crisis. Fanning out ratings to a broader category of very similar videos could have significantly limited the potential harm, including by demoting them. The case highlights inefficiencies in Meta's current approach during armed conflicts, compounding concerns the Board has expressed previously.

It is concerning that with the CPP activated and additional resources allocated, Meta did not identify on its own initiative the clear engagement abuse signals from the page, and that it only investigated the accounts behind it in response to Board questions. Accurate enforcement of the behavior-based policies could have prevented the harms from these violating accounts, rather than relying on content-based downstream mitigations prone to a high failure rate.

The Oversight Board's Decision

The Board overturns Meta's decision to leave up the content without a High Risk AI label.

The Board recommends that Meta:

- Create a Community Standard for AI-generated content, separate from the Misinformation Community Standard, providing comprehensive rules on provenance preservation, AI labeling protocols and self-disclosure.
- Develop pathways for affixing High Risk and High Risk AI labels to content much more frequently, assisted by clearer escalation channels from automated systems and at-scale review, so that such labeling can occur at a significantly higher volume.
- Attach provenance information and invisible watermarks to content created by Meta AI tools, including applying Content Credentials (as laid out by the C2PA) at creation.
- Implement Content Credentials at scale and ensure they are clearly and consistently visible and accessible whenever the provenance details are available.

- Invest in stronger detection tools for AI-generated multi-format (audio, audio-visual and image) content.
- Publish a clear explanation of penalties for failure to self-disclose digitally created or altered content, including the criteria for penalties and consequent limitations.
- Amend the Misinformation Community Standard to ensure that swift review of misinformation that directly risks imminent violence or physical harm does not depend solely on signals from external partners. A CPP lever should allocate resources for timely, proactive detection of such violating content, supported by in-house expertise and action, including labeling and investigating posting accounts and pages.

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

Full Case Decision

1. Case Description and Background

On June 13, 2025, Israel launched a major air strike targeting Iranian nuclear and military facilities, among other sites. Israel's leaders said the attacks were aimed at preventing the growth of Iran's nuclear program. This triggered the two nations exchanging intense attacks for more than a week and a half. Iran fired hundreds of missiles at Israeli cities, many of which were intercepted by Israel's defenses, while Israel struck multiple sites across Iran, including near the capital of Tehran. On June 18, United Nations (UN) Secretary-General Antonio Guterres stated he was "profoundly alarmed" by the military escalation, adding that "any additional military interventions could have enormous consequences, not only for those involved but for the whole region and for international peace and security at large." On June 21, the United States carried out a series of attacks targeting nuclear sites in Iran. On June 24, a ceasefire was announced between Israel and Iran.

The 2025 Israel–Iran conflict signaled an inflection point, with the growing influence of generative AI content on social media becoming known as its own "soft war." The British Broadcasting Corporation reported that three deceptive AI-generated videos of the conflict garnered more than 100 million views. Israel's Minister of Foreign Affairs shared a video of an attack on Evin Prison in Tehran, which forensic analysis later deemed likely to be an AI-generated video, despite an attack on the prison having actually taken place. Research from the Citizen Lab at the University of Toronto reported on a coordinated network of inauthentic profiles on X (formerly Twitter), allegedly associated with Israel, encouraging Iranians to push back against their government. The Israeli government also reported bot-driven campaigns by Iran aimed at shaping opinions around the conflict and the impact of their attacks on Israel.

Deceptive AI-generated content has been a growing and persistent challenge in crises and conflicts throughout the globe in recent years. This challenge is exacerbated in places where freedom of expression is under pressure and crackdowns on independent media and internet shutdowns obstruct credible information that can debunk deceptive campaigns. During its deliberations on this case, the Board observed how the U.S. operation to capture the President of Venezuela and mass anti-government protests in Iran involved claims that deceptive AI-generated content was authentic and counterclaims that authentic content was fabricated. Both situations challenged the public's ability to distinguish fabrication from fact, emblematic of the liar's dividend, risking a general distrust of all information.

Several technical approaches have emerged to help platforms and users distinguish between synthetic or manipulated and authentic media. One approach is tracing provenance i.e., the verifiable history of a digital asset, such as an image, video or document. The Coalition for Content Provenance and Authenticity (C2PA) has set out technical standards to embed provenance information as metadata in content, which allows platforms to more easily identify AI-generated content and apply labels to inform users. While these tools are still evolving and none are a perfect solution, they are seen as "the floor when it comes to responsible generation and dissemination of AI-generated content." In parallel, investments in automated detection, such as

classifiers, may provide a route to discover other signals that content has been AI-generated.

This case is emblematic of these challenges. On June 15, as the Israel-Iran conflict escalated, a video was posted to a Facebook page that claimed to be a news source and had 161,000 followers. The video depicted extensive damage to buildings, surrounded by plumes of smoke and rubble, with overlaid text in English reading “Live now – Haifa Towards Down" [sic] with the posting date. The post presumably referred to Haifa, a city in northern Israel. The video seems to be very similar to one originating on TikTok and identified by independent fact-checkers Agence France-Presse (AFP) as false and AI-generated (AFP only provides stills of the rated video, but these are identical to frames in the case content). A caption on the Facebook post in English listed many headline-style phrases linked to the conflict as well as unrelated terms and hashtags, without a clear narrative. It mentioned ongoing conflict, global political leaders, wildfires, missiles and more. The post received more than 700,000 views, with several comments noting that the content was AI-generated.

Six users reported the case content a total of nine times, but Meta’s automated systems did not prioritize it for human review. On the same day the content was posted, a misinformation classifier enqueued it to third-party fact-checkers, but it was never reviewed or rated. This was not unusual, as Meta flags a significant volume of potential misinformation that exceeds fact-checkers' review capacity.

After exhausting internal appeal procedures within the company, one of the reporting users appealed Meta’s decision to leave the content up to the Board. After the Board selected this case, Meta confirmed the post did not violate the Misinformation Community Standard because it did not “directly contribute to the risk of imminent physical harm.” Meta also defended its decision not to label the content. It took no action against the page or accounts responsible for the content.

Because of obvious signals of deception around this post, the Board asked Meta a series of questions about the identity and behavior of the page and accounts behind it. This prompted an investigation, leading Meta to identify that the page administrators had violated rules on engagement abuse and inauthenticity. The company then

permanently disabled three accounts, which removed the page and the case content from the platform.

2. User Submissions

In their submission to the Board, the user who requested removal of the content complained that Meta allows "terrorism acts" on their platform. There was no clear indication in their statement that they understood the content was AI-generated or misleading.

3. Meta's Content Policies and Submissions

I. Meta's Content Policies

Misinformation Community Standard

Under the Misinformation Community Standard, Meta removes "misinformation or unverifiable rumors that expert partners have determined are likely to directly contribute to a risk of imminent violence or physical harm to people." In countries "experiencing a heightened risk of societal violence," Meta works "proactively with local partners to understand which false claims may directly contribute to a risk of imminent physical harm" to identify and remove content making those claims.

Under the subheading "manipulated media," Meta states that for content that does not otherwise violate the Community Standards, it may add an informative label to the post when it is 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." The Misinformation policy also requires users to disclose whenever they post "organic content with photorealistic video or realistic-sounding audio that was digitally created or altered." Failure to use the AI-disclosure tool may result in penalties.

Elsewhere, Meta also prohibits content and behavior that "often overlap with the spread of misinformation." This includes Community Standards on account integrity,

deceptive practices and coordinated inauthentic behavior. For all other misinformation that does not violate its Misinformation Community Standard, Meta focuses on "reducing its prevalence or creating an environment that fosters a productive dialogue." Outside of the U.S., Meta relies on independent third-party fact-checkers to review and rate content, which can result in labels corresponding to the rating being added to the content. Ratings include "false" and "altered" and can lead to reduced distribution of content. Meta uses technology to surface potential misinformation for fact-checkers to review, and fact-checkers can also identify on their own content to review. In January 2025, Meta announced that it was ending the third-party fact-checking program in the U.S. and moving towards a Community Notes model instead.

Meta's Partner Monetization Policies outline the rules for pages "earning money on the platforms" and note that content flagged as misinformation or clickbait may not be eligible for monetization. Content Monetization Policies further outline the rules for "creating brand safe, monetizable content" and restrict or reduce monetization on content that depicts or describes certain subjects, such as "tragedy and conflict, including property damage." Relatedly, the accounts linked to the page in this case were both removed due to account level violations around engagement abuse.

II. Meta's Submissions

Meta stated that the post did not violate the Misinformation policy requiring removal of content "likely to directly contribute to the risk of imminent physical harm." Their decision took into consideration that no independent expert, such as a local partner, flagged the content or any related misinformation trend to them.

Meta did not apply any label to the content under the manipulated media rules. Meta applies three different labels to manipulated media: AI Info, High Risk, or High Risk AI.

The **AI Info label** is applied automatically to content when Meta detects "industry standard AI image indicators or when people disclosed that they were uploading AI-generated content." As the Board revealed in the Alleged Audio Call to Rig Elections in Iraqi Kurdistan case, Meta is currently only able to automatically identify and put the AI Info label on static images, relying on metadata that many generative AI tools embed in

such content. Audio or video content requires self-disclosure from users for the label to be applied. There was no self-disclosure in this case. Under Meta's current processes, a label could not automatically be added in these circumstances.

The **High Risk label** applies to content that (i) creates a particularly high risk of materially deceiving the public on a matter of public importance; and (ii) has reliable indicators of being digitally created or altered. Unlike the AI Info label, the High Risk label is an escalation-only policy, meaning only Meta's in-house policy teams can apply the label after human review. There was no such escalation of the content in question prior to the Board's selection of this case.

The **High Risk AI label** is applied when content meets all the requirements of the High Risk label and has reliable indicators of being created or altered with AI. It is also an escalation-only policy, requiring human review. Meta's in-house policy teams examined this content only after the Board selected it. They determined that too much time had elapsed by then for a label to be relevant or urgent.

Meta considers available external and internal sources it deems credible when determining whether content is the product of AI or digitally created or altered. External sources may include news or independent third-party fact-checking organizations that are able to provide a technical basis for their determination, such as a reference to an AI detection model or a conclusion from a forensic expert.

As explained in the Protest Footage Paired With Pro-Duterte Chants decision, manipulated media labels do not result in any automated demotion of content or its removal from recommendations. Instead, users who reshare content with these labels may be shown a pop-up, which may organically decrease reach. The pop-up's messaging will depend on whether the content was created with AI or not.

Meta has several systems outside of the U.S. to identify and address potential misinformation. Some systems only send content for third-party fact-checker review, while others both send the content for review and apply a temporary demotion while awaiting review. Whether content is only sent for review or has its reach reduced is

dependent on the type of system that flags it, as well as factors like country and language involved.

According to Meta, the Crisis Policy Protocol (CPP) was activated at the time of the Israel-Iran conflict. Israel had been designated under the CPP since the October 7, 2023, attacks in the country. Iran was designated at the beginning of the June 2025 conflict. Activating the protocol allows the company to deploy a set of levers designed to strengthen its crisis response and enable its teams to assess and mitigate the risk of imminent harm. During this crisis, it did not lead to any changes in the automated moderation systems, and the case content was reviewed using existing models and thresholds.

Meta designated the Israel-Iran conflict a "Trending Event" to better support third-party fact-checkers in identifying and debunking viral false claims related to the conflict, given the high risk of misinformation being propagated. By the time a ceasefire was in place, fact-checkers had published several fact-checks related to the conflict. Meta states that it has fact-checkers in Israel and the Philippines (where the posting user is based), but not in Iran.

After the Board requested that Meta investigate the behavior of this page and linked accounts, the company disabled the accounts of three different page administrators.

One administrator was disabled for violating the Authentic Identity Representation policy by "engag[ing] in identity misrepresentation to mislead or deceive others, evade enforcement or violate our Community Standards." The second account was disabled under the Account Integrity policy for being owned by the same person/entity as a disabled account. The third was disabled under the Spam policy for engagement abuse. The Spam policy broadly prohibits various deceptive, misleading or overwhelming practices to inauthentically increase engagement on posts. This led to the removal of the page and content posted by it. Prior to its removal, the page had been eligible to monetize through Meta's Stars program.

The Board asked questions on labeling, AI detection, fact-checking, CPP implementation, page and account-level behaviors, and more. Meta responded to all the questions.

4. Public Comments

The Board received six public comments that met the terms for submission. Four comments were submitted from Europe and two from the U.S. To read public comments submitted with consent to publish, click here.

The submissions covered the following themes: content moderation during conflict and crisis, the prevalence of AI-generated content and the rise of coordinated inauthentic behavior in armed conflict, the limitations of Meta's definition of "imminent physical harm," the significance of fact-checking during armed conflict, the standards and implementation of manipulated media labels, the importance of C2PA standards in detection, and more.

5. Oversight Board Analysis

The Board selected this case to examine Meta's policies and enforcement practices related to the sharing of deceptive AI-generated content on its platforms, particularly in the context of armed conflicts. This case falls within the Board's strategic priorities of Crisis and Conflict Situations and AI and Automation.

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

Content Rules

The Board finds that the case content did not require removal under the Misinformation Community Standard, but Meta should have applied a "High Risk AI" label to the content under its rules on manipulated media.

The post did not violate Meta's Misinformation policies requiring removal, as it was not likely to contribute to imminent violence or physical harm to people. The video deceptively exaggerated the impact of an Iranian assault on Israel and was posted close in time to missiles falling on Haifa. While this likely added to the distress of those deceived, it was not likely to contribute to violence from Israeli civilians against perceived adversaries or directly influence the Israeli government's response. There is no indication of violence between communities within Israel in response to deceptive depictions of the attacks.

Notwithstanding this conclusion, it is concerning that Meta did not provide any analysis of the potential risk of harm in its analysis of the content. Instead, it concluded that as no Trusted Partner had flagged the content to them, the content did not violate its rules. This is not an acceptable posture when many Trusted Partners are informing the Board that the company is less responsive to outreach and concerns, in part due to a significant reduction in capacities for Meta's internal teams. Meta should be capable of conducting such assessments of harm itself, rather than rely solely on partners reaching out to them during an armed conflict. This was a widely followed page, the content was viral, Meta's misinformation classifier flagged the content and several users reported it. Credible sources like AFP debunked a very similar video, putting Meta on notice to proactively review deceptive claims that could have caused harm. The activation of the CPP should have ensured the necessary resources were available for Meta to conduct these kinds of assessments itself and proactively reach out to partners for on-the-ground context if needed. In these circumstances, the content should have been escalated for review.

Had that happened, it would have been clear that the content posed a material risk of misleading the public on an important matter at a critical time, and a "High Risk AI" label would have then been applied. The page representing itself as a news outlet alongside text over the video claiming to be "Live Now" and from Haifa, portrayed this content as genuine footage of an ongoing armed conflict where civilian lives were at risk. This user, whose page was already misrepresenting itself as a credible news source, would not have revealed their deception by self-disclosing AI use. The unrealistic white birds flying through the video, and expert organizations like the AFP finding the content

to be AI-generated, should have caused Meta to assess whether labeling was needed. Meta should also have corrected its mistake as soon as the Board brought it to its attention. The Board's on-platform research revealed several versions of this video and related screenshots circulating across Meta's platforms weeks later, and the company had labeled none.

Had this specific post also been reviewed by third-party fact-checkers, this content likely would have received a false label and been demoted (based on AFP's rating of a very similar video). Meta's narrow approach to fanning out ratings to identical and near-identical content may have meant this content was not also rated (e.g., because of the addition of a text overlay to the video). Resource constraints and a remarkable volume of content make it difficult for fact-checkers to ensure timely review of all deceptive content, especially during a conflict or crisis. The Board reiterates that Meta should ensure that fact-checkers are adequately resourced and have guidance on prioritizing content from conflicts, to perform the challenging work that Meta counts on them to provide (see Protest Footage Paired With Pro-Duterte Chants).

It is concerning that during this crisis, with the CPP activated and additional resources allocated, Meta did not identify on its own initiative the clear engagement abuse signals from the page, and that it only investigated the accounts behind it in response to Board questions. Accurate enforcement of the behavior-based integrity and authenticity policies could have prevented the harms from this deceptive content backed by several violating accounts, rather than relying on content-based downstream mitigations prone to a high failure rate.

5.2 Compliance With Meta's Human Rights Responsibilities

The Board finds that, under Meta's human rights responsibilities, a manipulated media "High Risk AI" label should have been applied to the content, and Meta must do more to address the proliferation of deceptive AI-generated content on its platforms, including by inauthentic or abusive networks of accounts and pages.

Freedom of Expression (Article 19 ICCPR)

Article 19 of the International Covenant on Civil and Political Rights (ICCPR) provides for broad protection of expression, including political speech. This right includes the “freedom to seek, receive and impart information and ideas of all kinds" (Article 19, para. 2). 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).

Additionally, the UN Working Group on Business and Human Rights has expressed that “because the risk of gross human rights abuses is heightened in conflict affected areas,” due diligence by business should be “heightened accordingly” (A/75/212, para. 13). In a 2024 report on the Israel-Gaza conflict, the UN Special Rapporteur on freedom of expression and opinion surfaced that platforms are consistently failing to meet this responsibility in conflicts and noted the heightened risks from dis- and misinformation in such situations (A/79/319, para. 60, 66). In the Israel-Iran conflict, internet blackouts deeply impacted civilians’ access to information, creating a vacuum which deceptive AI-generated media rapidly filled (see PC-31545 WITNESS).

There is a robust suite of tools at Meta’s disposal to mitigate the potential harms of deceptive AI-generated content on its platforms. This case demonstrates that detection and labeling should be implemented more consistently, frequently and effectively to avoid likely near-term harms to users, especially in conflict situations where the stakes are much higher. This should be supported by adequate resourcing of third-party fact-checkers and guidance for prioritizing content from conflicts, as well as investment in accurate enforcement against account and page-level behavioral abuses.

I. Legality (Clarity and Accessibility of the Rules)

The principle of legality requires rules burdening 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.

The Misinformation Community Standard should provide more clarity to users and those enforcing the rules.

The general explanation of Meta collaborating with third parties to identify deceptive trends does not make clear that enforcement is entirely dependent on partners communicating concerns to Meta. Rather than clarify this, a different approach is warranted, for reasons outlined below.

The only detailed public description of the three manipulated media labels Meta uses is in the Board's decisions. The Board reiterates its recommendation that Meta fully describe the three manipulated media labels it applies, the criteria for applying them and their consequences (Protest Footage Paired With Pro-Duterte Chants), while noting this approach must also evolve.

Furthermore, the Manipulated Media policy does not publicly describe the penalties that "may be" applied if a user fails to self-disclose AI use. Meta explained to the Board that these penalties are only applied on escalation in response to repeated failures and may affect content distribution or temporarily result in suspension of certain account features. Meta exercises seemingly broad discretion in this area and clearer information should be provided to users.

Presenting more information on content provenance, labeling and self-disclosure in the Misinformation Community Standard could create confusion by conflating efforts to mitigate deception with positive measures to promote information integrity. Not all use of AI-generated content, and not all applications of labels, will be responsive to attempted deception. Outlining these rules in a separate Community Standard could help clarify Meta's approach and improve user behavior.

II. Legitimate Aim

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

In the Altered Video of President Biden case, the Board emphasized that "preventing people from being misled is not, in and of itself, a legitimate reason to restrict freedom of expression." However, the Misinformation Community Standard also aims to mitigate the risk of imminent physical harm or violence to people, which is a legitimate aim with respect to the rights of others (General Comment 34, para. 28).

III. 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).

The UN Special Rapporteur on freedom of expression has stated that "during armed conflict, people are at their most vulnerable and in the greatest need of accurate, trustworthy information to ensure their own safety and well-being. Yet, it is precisely in those situations that their freedom of opinion and expression [...] is most constrained by the circumstances of war and the actions of the parties to the conflict and other actors to manipulate and restrict information for political, military and strategic objectives" (A/77/288, para. 1).

The Special Rapporteur has also emphasized that "companies have tools to deal with content in human rights-compliant ways, in some respects a broader range of tools than those enjoyed by states. This range of options enables them to tailor their responses to specific problematic content, according to its severity and other factors" (A/74/486, para. 51).

In assessing the necessity and proportionality of potential measures, the Board considered the following: (a) that the page misrepresented itself as a credible news source; (b) that the content directly related to an ongoing armed conflict; (c) the vulnerability of civilians seeking verified information in the midst of that conflict; (d) the well-documented rapid spread of AI-generated deceptive content during this conflict (see PC-31528 Alan Turing Institute); (e) the cross-platform distribution of similar or near-identical content; and (f) the engagement and monetization incentives of generating manipulated media during conflicts.

The Board finds that placing a "High Risk AI" manipulated media label on the content would meet the requirements of necessity and proportionality and is concerned by Meta's failure to do so. This would be much less intrusive than removal, given that the deception was unlikely to lead to imminent harm. Currently, such labeling would be informational and would not result in demotion or removal from recommendations. Meta would show a pop-up to users who try to reshare content with this label. A label would help reduce the impact of this deception on users seeking accurate information online about the conflict.

Previous Board cases assert that the spread of manipulated media without a label may erode trust in the authenticity of content on the platform more broadly. This is particularly true during armed conflict, where manipulated media depicting violations of international humanitarian law could diminish confidence in these legal frameworks and the protections to civilians they provide. To meet its human rights responsibilities, Meta should have escalated review of the content without relying on external third parties, to label the content without delay.

The current mechanisms for affixing even the standard label of AI Info to video (user self-disclosure or an escalation to the Content Policy team) are neither robust nor

comprehensive enough to contend with the scale and velocity of AI-generated content, particularly during a crisis or conflict where there is heightened engagement on the platform. The Board notes that a system overly dependent on self-disclosure of AI usage and escalated review (which occurs infrequently) to properly label this content cannot meet the challenges posed in the current environment.

Some Members noted Meta's definition of "imminent physical harm or violence" is not inclusive of the various ways deceptive AI-generated content could have less direct but still serious societal impacts during an armed conflict. It could, for example, undermine access to credible information needed to hold government actors and other parties to account, increasing populations' vulnerability to manipulation, and enabling other forms of deceptive influence. For these Board Members, "manipulated media" labeling that is not coupled with demotion or removal from recommendations is insufficient to address those concerns. Suppressing the reach of such content during high-risk moments like armed conflicts would be a necessary and proportionate intervention. The Board acknowledges that a false rating from third-party fact-checkers on this content would have also resulted in this outcome.

Industry-wide, there are significant technical challenges to ensure the accurate and consistent labeling of AI-generated content. Meta explained to the Board that it uses industry standard indicators – being the metadata that generative AI tools often embed in content – to automatically label static images. However, not all generative AI tools currently attach the metadata necessary to apply a label. Even if a tool does attach the metadata, users can easily strip it from the content before sharing it on social media. Meta's limited mechanisms may reflect these current challenges; however, it is the responsibility of the company to proactively respond to rapidly evolving technology. This is an issue that extends beyond deceptive content, as these limitations obscure users' ability to verify the authenticity of all information. These obstacles will only intensify as the volume and quality of AI-generated video and audio continues to outpace available detection and labeling tools. The Board strongly encourages Meta to prioritize refining its detection and labeling mechanisms to better capture all forms of AI-generated content on its platforms, to properly inform users when they may be

interacting with manipulated media, and to ensure focus on those content types that pose the highest risks.

The Board recognizes Meta's presence on the steering committee of the C2PA. C2PA states that as information sharing rapidly transforms, it is critical to trace the provenance of media. It provides an open technical standard to establish the origin of and edits to digital content. Reports that Meta is not consistently implementing C2PA standards – even on AI-generated content from its own tooling – are concerning. A recent study showed that, when inspected by C2PA's tools, only a portion of the images and video generated by Meta's AI tools provided Content Credentials and received proper labeling.

The case content appears to have first originated off Meta's platform on TikTok, before being quickly reshared across platforms, with similar posts appearing on Facebook, Instagram and X, despite the report from AFP on the falsity of similar content. Public comments emphasized the need for strong cross-platform collaboration during periods of armed conflict to reduce and mitigate the pace of spread of deceptive AI-generated content.

As outlined above, the CPP and Trending Events designations should have allowed Meta to ensure more effective support to third-party fact-checkers during this crisis. In particular, fanning out ratings to a broader category of very similar videos could have significantly limited the potential harm, including by demoting it. The case highlights inefficiencies in Meta's current approach during armed conflicts, compounding concerns the Board has expressed in previous cases in different contexts (Protest Footage Paired With Pro-Duterte Chants decision).

6. The Oversight Board's Decision

The Board overturns Meta's decision to leave up the content without a High Risk AI label.

7. Recommendations

A. Content Policy

Misinformation

1. To ensure the swift review of misinformation that leads to risks of imminent physical harm or violence in crises, Meta should amend the Misinformation Community Standard to ensure that enforcement of this rule does not depend on signals from external partners. There should be a lever under the Crisis Policy Protocol to allocate resources for timely, proactive detection of such violating content, supported by in-house expertise, to identify, review and action content under the policy (including affixing labels under the Manipulated Media policy and investigating posting accounts and pages that show signals of engagement abuse).

   The Board will consider this implemented when Meta updates its Misinformation policy to reflect these requirements for the Physical Harm or Violence category.

AI-Generated Content

2. To help advance trust in information on Meta's platforms, Meta should create a Community Standard for AI-generated content, separate from the Misinformation Community Standard. The new Community Standard should provide comprehensive details on provenance preservation (i.e., capturing the detailed facts about the history of a piece of digital content), AI labeling protocols and self-disclosure rules.

   The Board will consider this implemented when Meta publishes a new Community Standard specifically on AI-generated content.

3. To improve the clarity of its rules, Meta should publish a clear explanation of penalties for failure to self-disclose digitally created or altered content. It should provide criteria for penalties and list which account features are consequently limited and for how long.

The Board will consider this implemented when Meta updates the Community Standard to include these penalty details and makes the revised guidance available in its public Transparency Center.

B. Enforcement

Provenance

4. To ensure users can reliably identify AI-generated content, Meta should implement Content Credentials (as laid out by the Coalition for Content Provenance and Authenticity) at scale and ensure that they are clearly and consistently visible and accessible to users whenever the provenance details are available. Provenance should not remain solely internally detectable or limited to back-end systems.

   The Board will consider this implemented when Meta provides a report explaining the changes it made to its interfaces and products to ensure that Content Credentials are consistently and clearly shown to users when available.

   Detection and labeling

5. To improve detection and labeling accuracy, Meta should invest in stronger detection tools for AI-generated multi-format (audio, audio-visual and image) content. Tooling should support escalation teams to better identify generative AI content trends, including potential harms around deceptive content in crisis situations.

   The Board will consider this recommendation implemented when the company confirms that stronger tools have been adopted and shares transparency data on the performance of these tools. These findings must be disaggregated by language and country, and whether the Crisis Policy Protocol was activated. This data must reflect comparable periods of time before and after the introduction of these changes.

6. To ensure more accurate labeling, Meta should attach provenance information and invisible watermarks to content created by Meta AI tools, so it can be consistently detected and labeled across platforms. This should include applying Content Credentials at the point of creation alongside using industry standard indicators for attribution to all content generated by Meta AI.

   The Board will consider this recommendation implemented when the company shares with the Board a report on how consistently Meta AI attaches and preserves provenance data and invisible watermarks to content shared on the platform.

7. To make the use of High Risk and High Risk AI labels on deceptive content more consistent, Meta should develop pathways for affixing those labels to content much more frequently, assisted by clearer escalation channels from automated systems and at-scale review, so that such labeling can occur at a significantly higher volume.

   The Board will consider this recommendation implemented when there are new pathways for escalations to affix High Risk and High Risk AI labels to content, and Meta reports to the Board on the volume of these labels attached in 2026, by quarter. The absence of a denominator (i.e., the total volume of unlabeled content that does not meet this threshold) should not be a barrier to providing this information to the Board.

***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.