Altered Video of President Biden

2023-029-FB-UA

Summary

The Oversight Board has upheld Meta's decision to leave up a video that was edited to make it appear as though U.S. President Joe Biden is inappropriately touching his adult granddaughter's chest, and which is accompanied by a caption describing him as a "pedophile." The Facebook post does not violate Meta's Manipulated Media policy, which applies only to video created through artificial intelligence (AI) and only to content showing people saying things they did not say. Since the video in this post was not altered using AI and it shows President Biden doing something he did not do (not something he didn't say), it does not violate the existing policy. Additionally, the alteration of this video clip is obvious and therefore unlikely to mislead the "average user" of its authenticity, which, according to Meta, is a key characteristic of manipulated media. Nevertheless, the Board is concerned about the Manipulated Media policy in its current form, finding it to be incoherent, lacking in persuasive justification and inappropriately focused on how content has been created, rather than on which specific harms it aims to prevent (for example, to electoral processes). Meta should reconsider this policy quickly, given the number of elections in 2024.

About the Case

In May 2023, a Facebook user posted a seven-second video clip, based on actual footage of President Biden, taken in October 2022, when he went to vote in person during the U.S. midterm elections. In the original footage, he exchanged "I Voted" stickers with his adult granddaughter, a first-time voter, placing the sticker above her chest, according to her instruction, and then kissing her on the cheek. In the video clip, posted just over six months later, the footage has been altered so that it loops, repeating the moment when the president's hand made contact with his granddaughter's chest to make it look like he is inappropriately touching her. The soundtrack to the altered video includes the lyric "Girls rub on your titties" from the song "Simon Says" by Pharoahe Monch, while the post's caption states that President Biden is a "sick pedophile" and describes the people who voted for him as "mentally unwell." Other posts containing the same altered video clip, but not the same soundtrack or caption, went viral in January 2023.


A different user reported the post to Meta as hate speech, but this was automatically closed by the company without any review. They then appealed this decision to Meta, which resulted in a human reviewer deciding the content was not a violation and leaving the post up. Finally, they appealed to the Board.

Key Findings

The Board agrees with Meta that the content does not violate the company's Manipulated Media policy because the clip does not show President Biden saying words he did not say, and it was not altered through AI. The current policy only prohibits edited videos showing people saying words they did not say (there is no prohibition covering individuals doing something they did not do) and only applies to video created through AI. According to Meta, a key characteristic of "manipulated media" is that it could mislead the "average" user to believe it is authentic and unaltered. In this case, the looping of one scene in the video is an obvious alteration.

Nevertheless, the Board finds that Meta's Manipulated Media policy is lacking in persuasive justification, is incoherent and confusing to users, and fails to clearly specify the harms it is seeking to prevent. In short, the policy should be reconsidered.

The policy's application to only video content, content altered or generated by AI, and content that makes people appear to say words they did not say is too narrow. Meta should extend the policy to cover audio as well as to content that shows people doing things they did not do. The Board is also unconvinced of the logic of making these rules dependent on the technical measures used to create content. Experts the Board consulted, and public comments, broadly agreed on the fact that non-AI-altered content is prevalent and not necessarily any less misleading; for example, most phones have features to edit content. Therefore, the policy should not treat "deep fakes" differently to content altered in other ways (for example, "cheap fakes").

The Board acknowledges that Meta may put in place necessary and proportionate measures to prevent offline harms caused by manipulated media, including protecting the right to vote and participate in the conduct of public affairs. However, the current policy does not clearly specify the harms it is seeking to prevent. Meta needs to provide greater clarity on what those harms are and needs to make revisions quickly, given the record number of elections in 2024.

At present, the policy also raises legality concerns. Currently, Meta publishes this policy in two places: as a standalone policy and as part of the Misinformation Community Standard.

 There are differences between the two in their rationale and exact operational wording. These need to be clarified and any errors corrected.

At the same time, the Board believes that in most cases Meta could prevent the harm to users caused by being misled about the authenticity of audio or audiovisual content through less restrictive means than removal of content. For example, the company could attach labels to misleading content to inform users that it has been significantly altered, providing context on its authenticity. Meta already uses labels as part of its third-party fact-checking program, but if such a measure were introduced to enforce this policy, it should be carried out without reliance on third-party fact-checkers and across the platform.

The Oversight Board's Decision

The Oversight Board has upheld Meta's decision to leave up the post.

The Board recommends that Meta:

- Reconsider the scope of its Manipulated Media policy to cover audio and audiovisual content, content showing people doing things they did not do (as well as saying things they did not say) and content regardless of how it was created or altered.
- Clearly define in a single unified Manipulated Media policy the harms it aims to prevent – beyond users being misled – such as preventing interference with the right to vote and to participate in the conduct of public affairs.
- Stop removing manipulated media when no other policy violation is present and instead apply a label indicating the content is significantly altered and could mislead. Such a label should be attached to the media (for example, at the bottom of a video) rather than the entire post and be applied to all identical instances of that media on Meta's platforms.

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

Full Case Decision

1. Decision Summary

1. The Oversight Board upholds Meta's decision to leave up a video that was edited to make it appear as though U.S. President Joe Biden is touching his adult granddaughter's chest and which is accompanied by a caption accusing him of being "a pedophile." The Board

 agrees with Meta that the post does not violate Facebook's Manipulated Media policy as currently formulated, for two reasons: (1) the policy prohibits the display of manipulated videos that portray people saying things they do not say, but does not prohibit posts depicting an individual doing something they did not do; and (2) the policy only applies to video created through artificial intelligence (AI). Because the video does not show President Biden saying words he did not say and the clip was not altered using artificial intelligence, it does not violate the company's Manipulated Media policy. Additionally, the accusation in the caption does not violate the Bullying and Harassment policy. Leaving the post on the platform also aligns with Meta's human-rights responsibilities, which include protecting the right to vote and take part in public affairs.

2. Although the decision not to remove this post was consistent with Meta's human-rights responsibilities, the lines drawn by the policy more broadly lack persuasive justification and should be reconsidered. In its current form, the Manipulated Media policy fails to clearly specify the harms it is seeking to prevent, and the scope of its prohibitions is incoherent in both policy and technical terms. The policy's application to (i) only video content; (ii) only content generated or altered by AI; and (iii) content that makes people appear to say "words they did not say," is too narrow to meet any conceivable objective. It is inappropriately focused on the medium of communication and method of content creation rather than on preventing specific harms that may result from speech (e.g., to electoral processes). At the same time, Meta's primary reliance on removing violating content may lead to disproportionate restrictions on freedom of expression. The Board recommends less severe measures be considered.

3. The technology involved in the creation and identification of content through AI is rapidly changing, making content moderation in this area challenging. It is all the more challenging because some forms of media alteration may even enhance the value of content to the audience. Some media is manipulated for purposes of humor, parody or satire. The Board previously emphasized the importance of protecting satirical speech (see "Two Buttons" Meme decision).

4. The Board therefore recommends that Meta revise its Manipulated Media policy to more clearly specify the harms it seeks to prevent. Given the record number of elections taking place in 2024, the Board recommends that Meta embark on such revisions expeditiously.

This is essential because misleading video or audio in themselves are not always objectionable, absent a direct connection to potential offline harm. Such harms may include (but are not limited to) those resulting from invasion of privacy, incitement to violence, intensification of hate speech, bullying, and – more pertinent to this case – misleading people about facts essential to their exercise of the right to vote and to take part in the conduct of public affairs, with resulting harm to the democratic process. Many of these harms are addressed by other Community Standards, which also apply to manipulated media. The Board is not suggesting that Meta *expand* the harms addressed by the Manipulated Media policy, but that it provides greater clarity on what those harms are.

5. In addition, the company should eliminate distinctions based on the form of expression, with no relation to harm. It should extend the policy to cover audio in addition to video, and all methods of manipulating media, not only those using AI. It should also include content depicting people doing things they did not do, in addition to the existing provision for things they did not say. Furthermore, Meta's enforcement approach should encompass the use of less restrictive measures to enforce the Manipulated Media policy, such as attaching labels to misleading content to explain that it has been significantly altered or generated by AI.

2. Case Description and Background

6. This case concerns a seven-second video clip posted in May 2023 on Facebook – almost six months after the midterm elections and 18 months before the 2024 presidential vote. The clip was based on actual footage of President Biden taken in October 2022, when he went to vote in person during the U.S. midterm elections accompanied by his adult granddaughter, a first-time voter. In the original footage of this occasion, President Biden and his granddaughter exchanged "I Voted" stickers. President Biden, following his granddaughter's instruction, placed a sticker above her chest, and then kissed her on the cheek. In the video clip posted to Facebook, the footage has been altered so that it loops, repeating the moment when President Biden's hand made contact with his granddaughter's chest so that it appears as though he is inappropriately touching her. The soundtrack to the altered video is a short excerpt of the song "Simon Says" by Pharoahe Monch, which includes the lyric "Girls rub on your titties," reinforcing the creator's imputation that the depicted act was sexualized. The caption to the video states that President Biden is "a sick pedophile" for "touch[ing] his granddaughter's breast!!!" and it

 also questions the people who voted for him, saying they are "mentally unwell." While other posts containing the same altered video clip but not the same soundtrack or caption went viral in January 2023, the content in this case, posted months later, had fewer than 30 views, and was not shared.

7. A user reported the post to Meta for violating the company's Hate Speech policy. That report was automatically closed without review and the content left up. This reporting user then appealed the decision to Meta. A human reviewer upheld the decision. The same user then appealed to the Board.

3. Oversight Board Authority and Scope

8. The Board has authority to review Meta's decision following an appeal from the person who previously reported content that was left up (Charter Article 2, Section 1; Bylaws Article 3, Section 1). The Board may uphold or overturn Meta's decision (Charter Article 3, Section 5), and this decision is binding on the company (Charter Article 4). Meta must also assess the feasibility of applying its decision in respect of identical content with parallel context (Charter Article 4). The Board's decisions may include non-binding recommendations that Meta must respond to (Charter Article 3, Section 4; Article 4). When Meta commits to act on recommendations, the Board monitors their implementation.

4. Sources of Authority and Guidance

9. The following standards and precedents informed the Board's analysis in this case:

I. Oversight Board Decisions

10. The most relevant previous Oversight Board decisions include:
    - Removal of COVID-19 Misinformation policy advisory opinion
    - Armenian Prisoners of War Video
    - Knin Cartoon
    - India Sexual Harassment Video
    - "Two Buttons" Meme
    - Armenians in Azerbaijan

II. Meta's Content Policies

11. Meta's Misinformation Community Standard explains that Meta will remove content only when it "is likely to directly contribute to the risk of imminent physical harm" or will "directly contribute to interference with the functioning of political processes," or in the case of "certain highly deceptive manipulated media."

12. The rules relating to interference in political processes focus only on "voter or census interference" (Section III of the Misinformation policy). In other words, this section of the Misinformation policy applies to information about the process of voting, and not about issues or candidates.

13. The rules relating to "highly deceptive manipulated media" are outlined under Section IV of the Misinformation policy and on a separate Manipulated Media policy page. According to the latter, Meta removes "videos that have been edited or synthesized, beyond adjustments for clarity or quality, in ways that are not apparent to an average person" by means of AI or machine learning, and "which would likely mislead an average person to believe" that the "subject of the video said words that they did not say." The policy rationale emphasizes that some altered media "could mislead." In Meta's Misinformation Community Standard, the prohibition of manipulated media is further justified by the rationale that such content "can go viral quickly and experts advise that false beliefs regarding manipulated media often cannot be corrected through further discourse." According to the standalone Manipulated Media policy page, there is a policy exception for content that is parody or satire.

14. For misinformation that is not removed for violating Meta's misinformation policy, Meta focuses on "reducing its prevalence or creating an environment that fosters a productive dialogue." For the latter, Meta attempts to direct users to "authoritative information." Meta states that as part of that effort, it partners with third-party fact-checking organizations to "review and rate the accuracy of the most viral content on our platforms." This is linked to a detailed explanation of the Fact Checking Program. Third-party fact-checkers have a variety of rating options, including "false," "altered," "partly false" and "missing context." The rating "altered" is applied to "image, audio or video content that has been edited or synthesized beyond adjustments for clarity or quality, in ways that could mislead people." This is not limited to AI-generated content or content depicting a person saying something

they did not say. Fact-checking does not apply to statements politicians make, within or outside of election periods. Meta does not control the ratings its fact-checkers apply and it is outside of the Board's scope to receive appeals on the decisions of fact-checkers.

15. Based on the ratings that fact checkers give, Meta may add labels to the content. Content labeled "false" and "altered" is obscured by a warning screen, requiring the user to click through to see the content. Meta explained that the "altered" label obscures the content with a full screen overlay informing the user that "[i]ndependent fact-checkers say this information could mislead people" as well as a "see why" button. Meta told the Board it provides users with a link to an article providing background, which is authored by the fact-checker whose rating was applied to the content. Again, this is neither reviewed by Meta nor appealable to the Board. A user who clicks the "see why" button is given the option of clicking through to the third-party fact-checking article that explains the basis for the rating. When a piece of content is labeled "false" and "altered" by fact-checkers, Meta also demotes the content, meaning that it ranks lower in users' feeds.

16. Meta's Bullying and Harassment Community Standard prohibits various forms of abuse directed against individuals. It does not apply, however, to criminal allegations against adults, even if they contain expressions of contempt or disgust. Nor does it prohibit negative character or ability claims or expressions of contempt or disgust directed towards adult public figures because these types of statements can be a part of important political and social discourse.

17. The Board's analysis was also informed by Meta's value of voice, which the company describes as "paramount," and its value of authenticity.

III. Meta's Human-Rights Responsibilities

18. The UN Guiding Principles on Business and Human Rights (UNGPs), endorsed by the UN Human Rights Council in 2011, establish a voluntary framework for the human-rights responsibilities of private businesses. In 2021, Meta announced its Corporate Human Rights Policy, in which it reaffirmed its commitment to respecting human rights in accordance with the UNGPs. The Board's analysis of Meta's human-rights responsibilities in this case was informed by the following international standards:

- The rights to freedom of opinion and expression: Article 19, International Covenant on Civil and Political Rights (ICCPR), General Comment No. 34, Human Rights Committee, 2011; UN Special Rapporteur on freedom of opinion and expression, reports: A/HRC/38/35 (2018), A/74/486 (2019), A/HRC/44/49 (2020) and A/HRC/47/25(2021).
- The right to take part in the conduct of public affairs and the right to vote: Article 25, ICCPR, General Comment No. 25, Human Rights Committee, 1996.

5. User Submissions

19. The user who appealed this case to the Board stated the content was a "blatantly manipulated video to suggest that Biden is a pedophile."

20. The author of the post was notified of the Board's review and provided with an opportunity to submit a statement to the Board, but declined.

6. Meta's Submissions

21. Meta informed the Board that the content does not violate its Manipulated Media policy because the video neither depicts President Biden saying something he did not say nor is it the product of AI or machine learning in such a way that it merges, combines, replaces or includes superimposed content.

22. Meta explained to the Board that in determining whether media would "likely mislead an average person," it considers factors such as whether any edits in a video are apparent (e.g., whether there are unnatural facial movements or odd pixelation when someone's head turns, or mouth movements out of sync with the audio). It also reviews captions for videos to see whether any disclaimers are included (e.g., "This video was created using AI"). Furthermore, the company will assume a video is unlikely to mislead when it is clearly parody or satire, or involves people doing unrealistic, absurd or impossible things (e.g., a person surfing on the moon). However, it would remove a video (or image) that depicts actions or events should it violate other Community Standards, whether generated by AI or not.

23. Following a policy development process in 2023, Meta plans to update the Manipulated Media policy to respond to the evolution of new and increasingly realistic AI. Meta is collaborating with other companies and experts through forums such as the Partnership on AI to develop common industry standards for identifying AI-generated content in order to provide users with information when they encounter this type of media.

24. Meta explained that the content in this case was not reviewed by independent fact-checkers. The company uses a ranking algorithm to prioritize content for fact-checking, with virality a factor that would lead to content being prioritized in the queue. The content in this case had no reactions or comments, only about 30 views, and was not shared; it therefore was not prioritized. Meta explained that other posts containing the same video (but with different captions) were reviewed by fact-checkers, but those reviews had no impact on this case due to the nature of fact-checkers' review, as described below.

25. Meta explained that its fact-checking enforcement systems take into account whether a fact-checker rated an entire post (e.g., a video shared with caption) or specific components of a post (e.g., only a video) to have false information. Meta then uses technology to identify and label identical and near-identical versions of the rated content across Facebook and Instagram. For example, when a fact-checker rates a whole post, Meta would apply a label only to posts that include identical and near-identical video and caption. If they had rated the video alone, Meta would label all identical and near-identical videos regardless of any caption. Rating and labeling components of posts, such as videos, independently from captions, can be more effective as this impacts more content on Meta's platforms. However, there may be meaningful differences in posts that share near-identical media, such as when an altered video is shared with a caption discussing its authenticity. Meta is therefore careful not to apply labels to all instances of media, such as an altered video, when the caption with which it is shared makes a meaningful difference. Many posts containing the same video as in this case were rated by fact-checkers, but those ratings were applied to the entire post (i.e., only those with matching video and caption), rather than to all posts that included the video (regardless of caption). As such, those ratings did not impact the video in this post.

26. Additionally, the reference to President Biden as a "sick pedophile" contained an expression of contempt and disgust (that he is "sick") in the context of a criminal allegation

 (that he is a “pedophile”). However, because Meta’s Bullying and Harassment policy does not apply to criminal allegations against adults, it concluded that the caption was not violating.

27. The Board asked Meta eight questions in writing. Questions related to the rationale underlying the Manipulated Media policy and its limited scope; the detection of manipulated media; Meta’s assessment of when media is “likely to mislead”; and Meta’s fact-checking program and which labels Meta applies to fact-checked content. All questions were answered.

7. Public Comments

28. The Oversight Board received 49 public comments relevant to this case: 35 were submitted from the United States and Canada, seven from Europe, three from the Middle East and North Africa, two from Central and South Asia, one from Latin America and Caribbean, and one from Asia Pacific and Oceania.

29. The submissions covered the following themes: the scope of Meta’s Manipulated Media policy; challenges of deciding which content has been altered and when content may mislead; the distinction between content generated or altered by AI and content altered by other means; the question of what harms manipulated media may cause and what impact it may have on elections; appropriate measures to moderate manipulated media; the challenges of moderating manipulated media at scale; and the impact of enforcement of manipulated media on freedom of expression.

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

31. In October 2023, as part of ongoing stakeholder engagement, the Board consulted with representatives of civil-society organizations, academics, inter-governmental organizations and other experts on the issue of manipulated media and elections. The insights shared during this meeting also informed the Board’s consideration of the issues in this case.

8. Oversight Board Analysis

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

33. This case was selected to examine whether Meta's Manipulated Media policy adequately addresses the potential harms of altered content, while ensuring that political expression is not unjustifiably suppressed. This issue is pertinent as the volume of manipulated media is expected to increase, in particular with continuing technological advances in the field of generative AI. The case falls within the Board's strategic priority of Elections and Civic Space.

8.1 Compliance With Meta's Content Policies

34. The Board agrees with Meta that the content does not violate Meta's Manipulated Media policy as currently formulated. The video clip does not show President Biden saying words he did not say and it was not altered using artificial intelligence (AI) to create an authentic-looking video.

35. Moreover, according to Meta's policy, a key characteristic of "manipulated media" is that it misleads the "average" user to believe the media is authentic and unaltered (see also PC-18036 - UCL Digital Speech Lab, discussing this under the term "blatancy"). In this case, the footage has been altered to loop, repeating the moment when President Biden's hand made contact with his granddaughter's breast, making it appear as if he was touching her inappropriately. The alteration of the video clip as such – looping the scene back and forth – is obvious. Users can easily see that content has been edited. The majority of the Board find that while the video may still be misleading, or intended to mislead about the event it depicts, this is not done through disguised alterations. A minority of the Board believe that the image may fall under the spirit of the policy as it still could "mislead an average user."

36. The Board understands that resources for third-party fact-checking are limited and the content's virality is an important factor in determining which content is prioritized for fact-checking. It is therefore reasonable that the post in this case was not prioritized due to its

limited potential reach, while other posts featuring an identical video were fact-checked on the basis of their reach.

37. The majority of the Board believe the caption that accompanies the video, which accuses President Biden of being a "sick pedophile," does not violate the Bullying and Harassment Community Standard. For public figures, Meta generally prohibits the forms of abuse listed under "Tier I: Universal protections for everyone" in the Bullying and Harassment Community Standard. This includes, "Attacks through derogatory terms related to sexual activity." Meta's policy specifically allows, however, claims including criminal allegations, even if they contain expressions of contempt or disgust. The majority find that the statement that Biden is a "sick pedophile" includes such an allegation and, as part of the discussion of a public figure, falls within that exception. The minority find that the claim that Biden is a "sick pedophile," when accompanied with a video that has been altered with the goal of presenting false evidence for the claim, does not constitute a criminal allegation but a malicious personal attack – and should therefore be removed under the Bullying and Harassment Community Standard.

8.2 Compliance with Meta's Human-Rights Responsibilities

Freedom of Expression (Article 19 ICCPR)

38. Article 19 para. 2 of the ICCPR provides broad protection for expression of "all kinds." The UN Human Rights Committee has highlighted that the value of expression is particularly high when it discusses political issues, candidates and elected representatives (General comment No. 34, para. 13). This includes expression that is "deeply offensive," insults public figures and opinions that may be erroneous (General comment No. 34, para. 11, 38, and 49).

39. 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). The Committee further states that the free communication of information and ideas about public and political issues between citizens, candidates and elected representatives is essential for the enjoyment of the right to take part in the conduct of public affairs and the right to vote, Article 25 ICCPR (General comment No. 25, para 25.)

40. When restrictions on expression are imposed by a state, they must meet the requirements of legality, legitimate aim and necessity and proportionality (Article 19, para. 3, ICCPR). These requirements are often referred to as the "three-part test." The Board uses this framework to interpret Meta's voluntary human-rights commitments, both in relation to the individual content decision under review and what this says about Meta's broader approach to content governance. As in previous cases (e.g., Armenians in Azerbaijan, Armenian prisoners of war video), the Board agrees with the UN Special Rapporteur on freedom of expression that, 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," (report A/74/486, para. 41). Nonetheless, the Board has also previously acknowledged that Meta can legitimately remove certain content because its human-rights responsibilities as a company differ from the human-rights obligations of states (see e.g., Knin Cartoon decision).

I. Legality (Clarity and Accessibility of the Rules)

41. Rules restricting expression should be clearly defined and communicated. Users should be able to predict the consequences of posting content on Facebook and Instagram. The UN Special Rapporteur on freedom of expression highlighted the need for "clarity and specificity" in content-moderation policies (A/HRC/38/35, para. 46).

42. Meta's Manipulated Media policy raises various concerns from a legality perspective. Meta publishes this policy in two different places (with no cross-reference link), as a self-standing policy and as part of the Misinformation Community Standard (Section IV). There are differences between the two in the rationale for restricting speech and their operative language. As a public comment pointed out, the wording of the policy is inaccurate (PC-18036 - UCL Digital Speech Lab). It states content will be removed if it "would likely mislead an average person to believe: (…)" that "(t)he video is the product of artificial intelligence or machine learning." This appears to be a typographical or formatting error because the opposite is presumably correct, that the average person could be misled *precisely because* it is not clear that content is AI generated or altered (as reflected in the Misinformation policy). This is confusing to users and requires correction. Furthermore, the self-standing policy states that it requires "additional information and/or context to enforce." The

Misinformation Community Standard does not include this statement. It only mentions that verification of facts requires partnering with third parties. In other cases, the Board learned that when a rule requires "additional information and/or context to enforce," it is only applied on-escalation, meaning by specialized teams within Meta and not by human reviewers enforcing the policy at scale (for previous cases engaging with Meta's escalation processes, see e.g., Armenian Prisoners of War Video, Knin Cartoon and India Sexual Harassment video). It would be useful for Meta to publicly clarify whether the Manipulated Media policy falls within this category or not.

II. Legitimate Aim

43. Restrictions on freedom of expression must pursue a legitimate aim (Article 19, para. 3, ICCPR), including to protect "the rights of others."

44. According to the policy rationale of the Manipulated Media policy, as presented in Section IV of the Misinformation Community Standard, it aims to prevent misleading content going viral quickly. Meta explains that "experts advise that false beliefs regarding manipulated media cannot be corrected through further discourse" without providing further evidence of this claim. The explanation in the standalone Manipulated Media policy is even less insightful, only stating that manipulated media could "mislead," without linking this to any specified harm.

45. Preventing people from being misled is not, in and of itself, a legitimate reason to restrict freedom of expression (General comment No. 34, para. 47 and 49). https://www.oversightboard.com/decision/FB-RZL57QHJThis is especially relevant in the context of political participation and voting, where contested arguments are an integral part of the public discourse (General comment No. 25, para 25) and competing claims may be characterized as misleading by some, and accurate by others. Additionally, media may be manipulated for purposes of humor, parody or satire and may as such constitute protected forms of speech (see "Two Buttons" Meme decision). In its submissions, Meta failed to explain, in terms of its human-rights responsibilities, what outcome the policy aims to achieve beyond preventing individuals being "misled" by content altered using AI (see also PC-18033 - Cato Institute). Meta did not explain whether it is consciously departing from international standards in adopting the Manipulated Media rule, per the

Special Rapporteur's guidance (report A/74/486, at para. 48 and report A/HRC/38/35, at para. 28).

46. Protecting the right to vote and to take part in the conduct of public affairs is a legitimate aim that Meta's Manipulated Media policy can legitimately pursue (Article 25, ICCPR). As public comments for this case illustrate, there is a broad range of views regarding how manipulated media can affect public trust in online information and in media more broadly and thus interfere with political processes. (See e.g., PC-18035 - Digital Rights Foundation; PC-18040 - Institute for Strategic Dialogue; PC-18045 - Tech Global Institute). Protecting the right to vote and to take part in the conduct of public affairs can justify taking measures against manipulated media, as long as Meta specifies the objectives of these measures and they are necessary and proportionate.

III. Necessity and Proportionality

47. Under ICCPR Article 19(3), necessity requires that restrictions on expression "must be appropriate to achieve their protective function." The removal of content would not meet the test of necessity "if the protection could be achieved in other ways that do not restrict freedom of expression," (General Comment No. 34, para. 33). Proportionality requires that any restriction "must be the least intrusive instrument amongst those which might achieve their protective function," (General Comment No. 34, para. 34). Social-media companies should consider a range of possible responses to problematic content beyond deletion to ensure restrictions are narrowly tailored (A/74/486, para. 51).

48. The Board acknowledges that Meta may put in place necessary and proportionate measures to prevent harms caused by manipulated media. Manipulation of media is often difficult for viewers to detect and may be especially impervious to normal human instincts of skepticism. Although humans have been aware for millennia that words may be lies, pictures and especially videos and audio impart a false veneer of credibility. While judgments about misinformation usually center on evidence for or against the propositions contained in a disputed message, judgments about manipulated media focus on the means by which the message was created. A central characteristic of "manipulated media" is that it misleads the user to believe that media is authentic and unaltered. This

 raises fewer risks that content moderation will itself be biased against particular viewpoints or misinformed.

49. In addition to defining the legitimate aim that the Manipulated Media policy pursues, however, Meta also needs to assure that the measures it chooses to enforce the policy with are necessary to achieve that goal. The Board believes that in most cases Meta could prevent harm to users caused by being misled about the authenticity of audio or audiovisual content through less restrictive means than removal. Rather than promote trust, content removal can sow distrust and fuel accusations of coverup and bias. For example, Meta could attach labels to misleading content to inform users that it was generated or significantly altered, providing context on its authenticity, without opining on its underlying substance. Labels could achieve this aim without the need for full warning screens that blur or otherwise obscure the content, requiring the user to click through to see it. This would mitigate against the risk of over-removals, given the challenges of accurately identifying content that is misleading (see e.g., PC-18041 - American Civil Liberties Union; PC-18033 - Cato Institute; PC-18044 - Initiative for Digital Public Infrastructure). Choosing the less restrictive measure of labeling rather than removing content would assure Meta's approach to enforcing its Manipulated Media policy is consistent with the necessity requirement. Restricting the enforcement of the Manipulated Media policy to labeling does not prevent Meta from removing information, including manipulated media, which misleads about the modalities of elections and which interferes with people's abilities to take part in the election process. Such information is removed under Meta's policy on "Voter or census interference" (Section III of the Misinformation policy), which also applies to manipulated media.

50. The Board notes that Meta already attaches labels to content under its third-party fact-checking program. Fact-checking, however, is dependent on the capacity of third-party fact-checkers, which is likely to be asymmetrical across languages and markets, and has no guarantee of genuine expertise or objectivity. The enforcement of an updated Manipulated Media policy through labeling may be more scalable. Labels could be attached to a post once identified as "manipulated" as per the definition in the Manipulated Media policy, independently from the context in which it is posted, across the platform and without reliance on third-party fact-checkers. The Board is concerned about Meta's practice of demoting content that third-party fact-checkers rate as "false" or

"altered" without informing users or providing appeal mechanisms. Demoting 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.)

51. The Board sees little sense in the choice to limit the Manipulated Media policy to cover only people saying things they did not say, while excluding content showing people doing things they did not do. Meta informs us that at the time of introducing the rule, videos involving speech were considered the most misleading and easiest to reliably detect. Whatever the merits of that judgment when it was made, the Board is skeptical that the rationale continues to apply, especially as methods for manipulating visual content beyond speech have and continue to develop, and become more accessible to content creators.

52. Second, it does not make sense to limit the application of the rule to video and exclude audio. Audio-only content can include fewer cues of inauthenticity and therefore be as or more misleading than video content. In principle, Meta's rules about manipulated media should apply to all media – video, audio and photographs. However, including photographs may significantly expand the scope of the policy and make it more difficult to enforce at scale. This may lead to inconsistent enforcement with detrimental effects. If Meta sought to label videos, audio and photographs but only captured a small portion, this could create the false impression that non-labeled content is inherently trustworthy. Furthermore, in the process leading up to the policy advisory opinion on the Removal of COVID-19 Misinformation, Meta presented evidence that the effectiveness of labelling diminishes over time, possibly due to over-exposure. To avoid diminishing the effectiveness of labels applied to manipulated audio and video, the Board, at this point, recommends not to include photographs in the proposed scope expansion. However, it encourages Meta to conduct further research into the effects of manipulated photographs and consider extending its Manipulated Media policy to photographs if warranted and if Meta can assure effective enforcement at scale.

53. Third, the Board is also unconvinced of the logic of making the Manipulated Media rule contingent on the technical measures used to create content. Experts the Board consulted, including at a dedicated roundtable, as well as public comments, almost unanimously agreed that the rule should be agnostic on the technical methods used (see e.g., PC-18036

 - UCL Digital Speech Lab). There was broad agreement that non-AI-altered content is, for now, more prevalent and is not necessarily less misleading; for example, most phones have features to edit content (see e.g., PC-18047 - Nexus Horizon). Moreover, it is not technically feasible, especially at scale, to distinguish AI-generated or altered content from content that is either authentic or manipulated using means other than AI. For these reasons, the policy should not distinguish the treatment of "deep fakes" from content altered in other ways (e.g., "cheap fakes" or "shallow fakes").

54. For the preceding reasons, the majority of the Board uphold Meta's decision to leave up the content. However, some Board Members believe the content should be removed even under the current standards, on the grounds that a false video presenting what might be misinterpreted as evidence of a serious crime is not protected speech, directly harms the integrity of the electoral process and is defamatory. According to the minority, such harms will not be prevented through less intrusive means.

9. Oversight Board Decision

55. The Oversight Board upholds Meta's decision to leave up the content, based on the Community Standards as they now exist.

10. Recommendations

A. Content policy

1. To address the harms posed by manipulated media, Meta should reconsider the scope of its Manipulated Media policy in three ways to cover: (1) audio and audiovisual content, (2) content showing people doing things they did not do (as well as saying things they did not say), and (3) content regardless of the method of creation or alteration.

   The Board will consider this recommendation implemented when the Manipulated Media policy reflects these changes.

2. To ensure its Manipulated Media policy pursues a legitimate aim, Meta must clearly define in a single unified policy the harms it aims to prevent beyond preventing users being

misled, such as preventing interference with the right to vote and to take part in the conduct of public affairs.

The Board will consider this recommendation implemented when Meta changes the Manipulated Media policy accordingly.

3. To ensure the Manipulated Media policy is proportionate, Meta should stop removing manipulated media when no other policy violation is present and instead apply a label indicating the content is significantly altered and may mislead. The label should be attached to the media (such as a label at the bottom of a video) rather than the entire post, and should be applied to all identical instances of that media on the platform.

   The Board will consider this recommendation implemented when Meta launches the new labels and provides data on how many times the labels have been applied within the first 90-day period after launch.

***Procedural note:**

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

For this case decision, independent research was commissioned on behalf of the Board. The Board was assisted by an independent research institute headquartered at the University of Gothenburg, which draws on a team of more than 50 social scientists on six continents, as well as more than 3,200 country experts from around the world. The Board was also assisted by Duco Advisors, an advisory firm focusing on the intersection of geopolitics, trust and safety, and technology. Memetica, an organization that engages in open-source research on social media trends, also provided analysis. Linguistic expertise was provided by Lionbridge Technologies, LLC, whose specialists are fluent in more than 350 languages and work from 5,000 cities across the world.