AI Video Faking UK Politician's Immigration Views

2026-034-FB-UA

Summary

The Oversight Board has found that a deepfake video of a UK local politician should be removed from Facebook because it contains hate speech against refugees, overturning Meta's decision to leave it up. The Board found Meta needs to do much more to effectively address such content, including by increasing penalties for sharing harmful deepfakes, to protect users from potential deception.

Why This Matters

The case surfaces monumental challenges facing societies that social media companies must address – the increasing realism of deepfakes depicting elected officials misrepresenting their views on critical issues and misleading the public, and misinformation that fuels hatred against minorities, including immigrants. Women politicians, journalists and activists are disproportionately targeted by deepfakes that misrepresent their views, including as part of harassment campaigns, and less prominent public figures such as local politicians and activists often lack the resources to respond effectively. The spread of deepfakes erodes trust in all public information and challenges the boundaries between what is protected political speech and what violates Meta's rules on misinformation, hate speech or bullying and harassment.

About the Case

In November 2025, a Facebook user posted a short video showing the likeness of a Labour Party councillor in Scotland. The video, which appears to be AI-generated,

portrays her saying: “Refugees are welcome here, even if they rape our women, because white people do that too.”

The portrayal is realistic, though on closer inspection appears synthetic: the audio is not fully synchronized to her facial movements, indicating that it is a fabrication. The video was part of an album containing a photograph of several named women at an anti-far-right protest, including the councillor. The album and video together received more than 5,000 views, over 50 comments, over 20 reactions and more than 10 shares.

Two individuals, including the depicted politician, reported the video for violating the Bullying and Harassment policy, but Meta’s systems did not prioritize the post for human review, and it remained on Facebook. Both users appealed to Meta, but the post was kept up and not reviewed by a human. One of the users, not the politician involved, then appealed to the Board. When the Board brought the case to Meta’s attention, the company concluded the video did not violate its Community Standards and that it did not merit an AI label.

Key Findings

The Board evaluated this case with reference to three of Meta’s policies – Hateful Conduct, Misinformation, and Bullying and Harassment. The Board finds that Meta should have removed the AI-generated video under its Hateful Conduct policy.

The majority of the Board finds that the video violated Meta’s Hateful Conduct rules specifically because it alleges serious criminality and predatory sexual behavior against refugees as an entire group, and not against some refugees as Meta claimed.

In ruling on the case, the Board also considered whether the video, had it not violated the Hateful Conduct rules, would have met the threshold for removal under the misinformation policy. A majority of the Board finds that removal would not have been

warranted for misinformation, but that it should have received a “High Risk AI” label, and in its analysis concludes that Meta’s rules on AI labeling of deepfakes are inadequate.

The majority finds that Meta needs more robust policies on deepfakes, including expanding the situations when “high risk” labels can be applied, more measures to reduce the spread of deceptive AI content, increasing the penalties for accounts that repeatedly share it, and more transparency on data around when AI labels are applied.

A minority of the Board disagrees that the video should be considered hate speech because it is directed at the alleged views of the politician, rather than at all or most refugees; and that because the video is political speech, the threshold for removal should be higher.

There were also minority views on the Board’s findings under the misinformation rules. One minority said that a “High Risk” label is too restrictive and that a lesser “AI Info” label should have been applied. Another, separate minority said that the gravity of harms, coupled with the fact that Meta does not in practice apply many AI labels, would justify removal.

The Board finds that the video would not require removal under the Bullying and Harassment rules because the elected official is a public figure and therefore is entitled to less protection than private individuals under the policy.

The Oversight Board’s Decision

The Board overturns Meta's decision to leave up the content, requiring it to be removed under the Hateful Conduct Community Standard. In considering this case, the Board also finds shortcomings in labeling policies under the Misinformation Community Standard and makes the following recommendations regarding that policy.

The Board recommends that Meta:

- Change its manipulated media policy to lower the threshold for "high risk" labels, making sure this is not a crisis or election integrity measure only.
- Increase friction for users viewing content labeled as AI-generated or "High Risk AI", such as using an interstitial (a screen requiring click-through to view content) instead of a label.
- Subject all "High Risk" labeled content to demonetization, demotion or removal from recommendations, with escalating penalties for accounts that repeatedly share "High Risk" labeled content.
- Ensure that the rules and examples related to voting or census interference are phrased as globally applicable and not limited to the U.S. context.
- Publicly disclose annual data in its Transparency Center on the number of times it applies the "High Risk" and "High Risk AI" labels to content, including the number of times these labels are applied to content from or about politicians.
- Ensure that all labels under the Misinformation policy ("AI Info," "High Risk" or "High Risk AI") can be seen on content in the Meta Content Library, and that content that has received these labels is searchable via a filter.

The Board also reiterates the importance of its previous recommendation in the AI Generated Video in Israel-Iran Conflict decision, calling for Meta to 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.

Full Case Decision

1. Case Description and Background

In November 2025, a Facebook user posted a short video showing the likeness of a Labour Party councillor in Scotland. The video, which appears to be AI-generated, portrays her saying: “Refugees are welcome here, even if they rape our women, because white people do that too.” The portrayal is realistic, though on closer inspection appears synthetic: the audio is not fully synchronized to her facial movements, indicating that it is a fabrication. The video was part of an album (the totality of which was not subject to review) containing a photograph of several named women at an anti-far-right protest, including the councillor. The album also contained an unrelated video of another protest that appears AI-generated but did not mention the councillor. The caption to the album notes the councillor campaigned against closing refugee accommodation in her town.

The album and video together received more than 5,000 views, over 50 comments, over 20 reactions and more than 10 shares. One comment questioned the veracity of the video and criticized the user for posting it. Another comment, which Meta removed the same day it was posted, suggested the depicted politician wanted to be raped.

Two individuals, including the depicted politician, reported the video of the councillor for violating the Bullying and Harassment policy, but Meta’s systems did not prioritize the post for human review, and it remained on Facebook. Both users appealed to Meta, but again the post was kept up and not reviewed by a human. One of the users, not the councillor depicted in the video, then appealed to the Board. When the Board brought the case to Meta’s attention, the company concluded the video did not violate its Community Standards and that it did not merit an AI label.

This case is illustrative of two related systemic challenges facing social media companies globally: increasingly pervasive, realistic and deceptive synthetic portrayals of politicians, sometimes referred to as "deepfakes," and disinformation fueling hate against minorities, including immigrants. The Board addressed the connections between AI-generated disinformation and hate speech in the Posts Supporting UK Riots and Criticism of EU Migration Policies and Immigrants decisions. It has also addressed synthetic audio of politicians in the Alleged Audio Call to Rig Elections in Iraqi Kurdistan decision and manipulated video in the Altered Video of President Biden decision.

The content in this case was posted several months in advance of local elections. Though the councillor's seat was not up for re-election until May of 2027, her party was engaged in contests up and down the country where immigration and the housing of asylum seekers was a prominent issue. In 2025, demonstrations against the housing of asylum seekers in hotels took place across the United Kingdom after an asylum seeker in the south-east county of Essex sexually assaulted a woman and a 14-year-old girl. Protests continued in 2026, including in Scotland, with claims that asylum seekers pose safety concerns for women, which some characterized as racist disinformation weaponizing concerns about violence against women to fuel anti-migrant hate.

2. User Submissions

In their appeal to the Board, the user who reported the content alleged that the video was AI-generated and misrepresented the politician's views and could pose a threat to the woman's safety. From their perspective, people should be free to express their own views, but this should not extend to creating fake videos of politicians making inflammatory statements they did not make.

In a statement to the Board, the depicted councilor said that the video was AI-generated, and the words attributed to her "run contrary to everything [she] stands for and significantly impugned [her] integrity." The adverse impacts of the video were,

according to her, significant and quite traumatic. She made multiple unsuccessful attempts to have it removed, including reporting the incident to Meta and law enforcement. Notwithstanding this, she informed the Board she has remained active in advocating for the rights of refugees in Scotland and campaigning against the spread of hatred in her community.

3. Meta's Content Policies and Submissions

Meta's submissions focused on three policy areas: Misinformation, Hateful Conduct, and Bullying and Harassment. The company maintained that no action was required against the content under any of these policies, leaving it up without a label.

Misinformation

The policy rationale for the Misinformation Community Standard states that Meta removes "content that is likely to directly contribute to interference with the functioning of political processes." One of the three categories of misinformation Meta removes relates to political processes and is titled "voter or census interference." Eight examples of prohibited voter or census interference follow, including misinformation about "the dates, locations, times and methods for voting, voter registration or census participation," and "whether a candidate is running or not." Two examples are specific to the United States. The policy rationale notes Meta collaborates with Trusted Partners, a global network of non-governmental organizations (NGOs), humanitarian agencies and human rights researchers that flag emerging risks from content on Meta's platforms, to assess the truth of content and whether it is "likely to directly contribute to the risk of imminent harm." The policy rationale also acknowledges that "people often use misinformation in harmless ways" such as "to exaggerate a point" or in "humor or satire."

Meta informed the Board it did not remove the video in this case because, according to the company, Trusted Partners did not flag it. Meta further concluded the content was not "likely to directly contribute to a risk of interference with the functioning of political processes" as it did not contain any prohibited "voter or census interference" claims.

For certain misinformation that does not violate the rules to warrant removal, Meta focuses on "reducing its prevalence or creating an environment that fosters a productive dialogue." The company does this by focusing on "providing users with helpful information when there is potentially misleading or confusing content," including through labels.

There are three types of labels:

- 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." It is less prominent than the other two labels, appearing at the top of a post.
- The **"High Risk" label** applies to content under the "manipulated media" policy 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 applied under Meta's escalation-only policy, meaning only Meta's in-house policy teams can apply the label after human review. It is more prominent, appearing at the bottom of a post, and links to a resource to learn more about digitally created or altered content.
- The **"High Risk AI" label** is also applied under the "manipulated media" policy 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 requires people to disclose, using in-app tooling, whenever they post organic content with "photorealistic video or realistic-sounding audio that was digitally created or altered." Meta states that failure to use the AI-disclosure tool "may result in penalties."

Meta said the user who created the post did not disclose AI use which is why it did not receive an "AI Info" label. The company informed the Board that it can apply this label automatically when it detects "industry standard AI image indicators," but that these were not present in this case. In response to Board questions, Meta confirmed that it did not penalize this user for failing to disclose AI use. Moreover, Meta's enforcement of the AI self-disclosure requirement has been limited to individual cases escalated to its in-house policy teams rather than applied at scale. Meta also noted it is now able to detect AI industry-shared signals for video and images, and is in the process of rolling this out across all devices (web and mobile) and surfaces (reels, feed, etc). Separately, they are developing classifiers capable of detecting AI-generated content at scale for the purpose of applying the less prominent "AI info" label, regardless of whether the videos contain industry standard provenance signals.

Meta confirmed that, in its view, the content also did not merit more prominent "high risk" labeling under its manipulated media rules. As no escalated review of the content occurred, an assessment to apply this label was not carried out. In addition to challenges in identifying whether the content was digitally altered or created using AI, Meta explained the content did not create a "particularly high risk of materially deceiving the public on a matter of public importance." This was because, in Meta's judgment, there was no urgency related to a time-sensitive event – such as an imminent election or crisis – that would elevate the risk of material deception: the closest elections were months away. Second, Meta noted that the content was satirical, inverting the politician's public position, and this understanding was reflected in at least one user comment. Third, the content received little engagement, reducing the likelihood the public at large would be materially deceived.

In response to Board questions, Meta explained its threshold for deploying the more prominent "High Risk" labels is exacting. It has only used the label in a very limited number of cases. Meta justifies its high threshold because: (a) adding such labels calls into question the authenticity of the content and can harm the reputation of users posting it; (b) AI detection is imperfect, and could result in unlabeled content looking more trustworthy when it is not; (c) it may have a chilling effect on freedom of expression and (d) extending such labels beyond critical events may invite growing demands from politicians, governments and advertising partners to label speech they dislike. Meta also noted that because content often does not have AI provenance signals, it relies on other factors, such as how the content "looks"; whether synthetic content is intended as humor also factors into whether they would find content to be deceptive or not. Meta said it wants to avoid being the "arbiter of truth" and prefers a higher threshold that prevents the company from having to make subjective assessments of what is deceptive or important.

Outside the U.S., Meta relies on independent third-party fact-checkers to review and rate content, which can also 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 content to review. However, Meta confirmed that in this case its automated systems did not send the content to third-party fact-checkers and they did not select it for review, most likely because of the low virality. Meta eliminated third-party fact-checking in the U.S. only, in favor of Community Notes, and the Board provided an opinion on the advisability of making a similar shift elsewhere in the world.

The posting user's account was not monetized in this case. Meta's Partner Monetization policies set the rules for pages the company pays for posting on its platforms based on engagement, and note that content flagged as misinformation is not eligible for

monetization. Content Monetization policies further restrict or reduce monetization for content about “debated social issues” that "exploits controversial political or social issues for commercial purposes,” including immigration, presented in a “polarizing or inflammatory manner.”

Hateful Conduct

Meta’s Hateful Conduct Community Standard prohibits “allegations of serious immorality and criminality” against people based on their protected characteristics or status as refugees. It lists comparing people to “sexual predators” and “violent criminals” as examples of violations.

Meta found the statement, “Refugees are welcome here, even if they rape our women because white people do that too” did not violate this rule, as it did not equate refugees with violent criminals or sexual predators based on who they are. Internal guidance to reviewers distinguishes between prohibited “generalizations” and permitted “behavioral statements.” Prohibited generalizations attribute inherent traits to an entire group (such as saying: [Group X] are terrorists; or [Group Y] are murderers). Permitted behavioral statements are assertions against the actions of members of a group (such as saying: [Group A] engages in terrorism; or [Group B] engages in murder). For Meta, the assertion in the video was about the actions of some refugees and did not claim that sexual violence is an inherent trait of refugees. It describes behavior attributed to members of a group, rather than equating a group’s identity with sexual criminality. The phrase “because white people do that too,” does not, in Meta’s view, change that meaning. Rather, it compares the behavior of two groups, attributing sexual violence to some white people also.

In previous cases, Meta has explained to the Board that behavioral statements may become violating where alleged severe criminality is explicitly attributed to all or most of the group. In the Criticism of EU Migration Policies and Immigrants case, Meta found

a statement referring to depictions of immigrants as “gang-rape specialists” was not violating, because it did not attribute the behavior to all or more than half of the group. Meta rejected the Board’s recommendation in that case to reverse its presumption that references to the behavior of a group should be read as referring to the whole or most of the group unless specified otherwise. Meta noted in its submissions to the present case that after rejecting this recommendation, it continues to presume statements like “refugees rape” do not mean “all” or “most” refugees and therefore do not violate its Hateful Conduct policy.

Bullying and Harassment

Meta’s Bullying and Harassment Community Standard distinguishes between public figures and private individuals to “allow discussion, which often includes critical commentary of people who are featured in the news or who have a large public audience.” Public figures are only protected from the most severe attacks. While Meta prohibits “unwanted manipulated imagery,” this rule only applies to content targeting private minors, private adults and minor involuntary public figures when they self-report. Meta defines public figures as “state- and national-level government officials,” candidates for those offices and other people "who receive substantial news coverage.” Meta determined that the politician depicted in the video is a public figure, and therefore not protected from “unwanted manipulated imagery.”

4. Public Comments

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

The submissions covered the following themes: the scale and prevalence of AI-generated content; dissemination of anti-immigration content; the use of satire, humor

and political expression on social media, including in AI-generated content; gendered harassment of women in public life; effectiveness of labeling as a response to misinformation; protection of speech against public figures in the Bullying and Harassment policy; misinformation policy gaps and political process interference; technical challenges in detecting the provenance of content; how misinformation and hate speech harm the information environment and democracy; and commercial and monetization incentives driving deceptive and hateful AI content.

In June 2026, the Board consulted with representatives of advocacy organizations, academics, inter-governmental organizations and other experts on the issue of political deepfakes, their impact on information and electoral integrity, and how platforms can prevent and mitigate harm while respecting freedom of expression.

Participants noted the growing prevalence of AI-generated content mimicking politicians and targeting hate at vulnerable communities. Research presented at the consultation documented that xenophobic AI-generated content in the UK received disproportionately higher engagement than comparable authentic content. Participants also noted that women politicians, journalists and activists are disproportionately targeted by deepfakes that misrepresent their views, including as part of harassment campaigns, and that less prominent public figures, like locally elected figures and activists, often lack institutional resources to respond to such attacks.

Participants raised concerns about Meta’s Misinformation policy being too narrow in the protections it provides to election integrity and democracy more broadly, only prohibiting a limited number of false claims about voting processes close in time to elections. They described this framework as structurally inadequate given that content posted on social media platforms persists indefinitely and is continuously circulated, sustaining harmful narratives well beyond any formal election period. Several participants described Meta's reliance on Trusted Partners to identify potential harm as

an unsatisfactory outsourcing of enforcement responsibility that systematically fails to deal with content with low initial engagement, especially in smaller markets. Finally, several participants called for more focus on AI-generated impersonation in content, as current policies seem to focus on rules on "impersonation" through inauthentic accounts (i.e., people setting up accounts to misrepresent that they are that person, rather than people sharing content to misrepresent the conduct of a person).

On satire, participants noted that the absence of any visible signals on AI-generated content depicting politicians left audiences without a reliable way to identify it as either fictional or real. One participant argued that deepfakes can be satirical, and that intent should be assessed in context, including by reference to the depicted person's public positions or actions. Several participants stressed that the more consequential harm of deepfakes lies not in whether individual viewers are deceived by any single piece of content, but in the erosion of trust in political information more broadly. When audiences are repeatedly exposed to realistic fabrications, they may lose confidence in their ability to distinguish them from authentic speech. Finally, participants noted that labeling has diminishing protective value in this environment. This happens because labeling is inconsistently enforced, but also because some people continue to react to content labeled as inauthentic as if it was authentic, especially when it reinforces preconceived biases.

Participants noted that a significant proportion of synthetic content is commercially rather than ideologically motivated, often produced by people not connected to the country the content concerns, who exploit Meta's monetization tools or other means of revenue-generation to profit from sensational – including hateful – narratives that engagement-based algorithms are likely to boost. Meta's creator incentive structures reward provocative content with higher engagement and algorithmic amplification, making the production of deceptive AI-generated content economically attractive even where individual posts are not directly monetized. Several participants observed that restricting monetization remains an underexplored tool for deterring deceptive AI use.

5. Oversight Board Analysis

The Board selected this case to examine how Meta's policies treat deceptive AI-generated content on its platforms, specifically when AI tools are used to impersonate politicians to fabricate their views on matters important to public discourse, including in ways that may constitute hate speech. The analysis assesses Meta's content policies and human rights responsibilities across three themes: misinformation; hate speech; and bullying and harassment.

The Board finds that while the content does not warrant removal under the Misinformation policy, it should have received a "High Risk AI" label; from a human rights perspective, the Board makes a number of recommendations aimed at improving how Meta responds to deceptive AI-generated content of politicians saying things they did not say on matters of public importance. Notwithstanding these findings, in this case, the Board found the content violated Meta's Hateful Conduct policy and should have been removed on this basis. Finally, the Board finds the content does not violate Meta's Bullying and Harassment policy.

Misinformation

The Board finds that the video of the depicted politician did not violate Meta's Misinformation Community Standard rules for removal, as it contained none of the prohibited misinformation claims for "voter or census interference."

For a majority of the Board, the Misinformation Community Standard rules on manipulated media required a prominent "High Risk AI" label to be added to the video because it was generated using AI and poses a particularly high risk of materially deceiving the public on a matter of public importance.

The video in this case has hallmarks of AI-generated media, but it would not be immediately obvious to most viewers that it is fake. Visible mismatches between voice and facial movement and unnatural facial expressions are cues of the video being AI generated but are only apparent on close inspection. The majority notes that Meta has informed the Board its classifiers are now capable of automatically detecting images and video at scale, even in the absence of provenance metadata, but at the time this content was posted, such detection would require human review. The video's realism, the undisclosed use of AI to completely fabricate (rather than alter or manipulate) footage, clearly distinguishes it from the type of "cheap-fake" editing seen in the Altered Video of President Biden case.

For a majority of the Board, the video seeks to deceive potential voters by presenting "evidence" of an elected politician's supposed views on the housing of refugees and public safety, a topic central to contemporary political discourse and high-profile protests across the country in question. Meta's submissions point to non-public criteria for imposing "High Risk AI" labels that mean these labels have, in practice, barely been applied at any meaningful scale. These non-public criteria establish an enforcement threshold that is exceedingly high given the public language of the rule, the neutral and non-denunciatory language of the label, and its limited enforcement effects (sanctions against users appear even less common than the labels themselves).

For these reasons, Meta's justification that there was no crisis or imminent election, and that the content had limited reach, is unconvincing. Neither factor means that the content was unlikely to deceive the public. Moreover, there are no signals of exaggeration or parody that would make it clear that it is satirical. The short length of the video, the depiction of the politician against a plain background without any markers of exaggeration or parody, in the context of an album where no other content is satirical or humorous, increases the risk of deception and points to likely intent to deceive. The Forum for Humor and Law, an academic project exploring how freedom of expression law protects humor, (see PC-32578) provided the Board useful examples of

satirical deepfakes, where an initial deception comes undone as the absurdity of the fabrication is realized (French politician Marine Le Pen in Islamic garb speaking Arabic; former Brazilian President Bolsonaro providing a COVID-era tutorial on handwashing). Based on the language of the Community Standards, a High-Risk AI label should have applied.

For the majority, Meta's human rights responsibilities require it to take a much more effective and robust response to the deceptive use of generative AI to fabricate a politician's words on matters of public importance than the company's current approach to "high risk" AI labelling.

For the majority, a combination of measures to better detect AI use and where it is deceptive on matters important to the public, add friction, reduce reach and penalize harmful deception through clearer labelling would be a necessary and proportionate restriction on freedom of expression. This would protect "the rights of others" to participate in public affairs and protect the politician from "unlawful attacks on [their] reputation" (Articles 17 and 25, International Covenant on Political and Civil Rights (ICCPR)).

The Board agrees that international standards on freedom of expression protect falsehoods, and that the value placed by the ICCPR on uninhibited expression concerning public figures in the political domain is particularly high (General Comment No. 34, at para. 38). As the Board observed in the Altered Video of President Biden case:

"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."

For a majority of the Board, there is a qualitative difference between a stated lie about what a politician said or did, and fabricated audiovisual "evidence" of the same lie. Such deepfakes can more easily convince people a falsehood is indisputable, in a way that risks harm to democratic discourse, public participation and the reputations of individuals. When the fabrication is highly realistic, purposefully deceitful and even malicious, it should not attract the greater protection international human rights law affords to political speech (General Comment No. 34, at para 34). As outlined above, there are no indications this video was satirical, which would make the falsehood more apparent and not harm the rights of others. The spread of generative AI technologies poses new challenges to the marketplace of ideas, including industrial level fraud that undermines democracy and people's rights to participate, and can enable lasting and widespread defamation of an individuals' reputation.

As it becomes increasingly challenging to distinguish manipulated or fabricated media from reality, synthetic and deliberately deceptive audiovisual representations of public officials pose a unique and serious risk to public trust in information, and in turn participation in democracy. Recent research shows that on average, people in the UK struggle to identify synthetic media depicting real people. The compounding impacts of such fakes can undermine mechanisms for holding politicians to account, not just at the ballot box but between elections (see: Preliminary Report of the United Nations Independent International Scientific Panel on AI, July 2026, at 2.7, from page 23). Specifically, the proliferation of falsified content can lead to a "liar's dividend," which is the advantage reaped by people who falsely claim that authentic evidence is fake or AI-generated to avoid accountability (Chesney and Citron, 2019). The United Nations Educational, Scientific and Cultural Organization (UNESCO) has described the broader danger of deepfakes as a "crisis of knowing," one that does not merely introduce falsehoods but erodes "the very mechanisms by which societies construct shared understanding." This may exacerbate pre-existing persuasion effects of platforms' engagement-based personalized feeds and recommendations, where manipulative

claims at-scale are paired with synthetic "proof" of those claims (*op. cit.*, at 3.5, from page 37). This challenge is global: one study identified 82 deepfakes targeting public figures across 38 countries in a year, with more than a quarter fabricating false statements from public figures (see Targets, Objectives, and Emerging Tactics of Political Deepfakes, Recorded Future, September 2024). As noted by NGO CEE Digital Democracy Watch in a public comment submitted to the Board, if left without action, deepfakes can contribute to "epistemic manipulation," distorting "the informational conditions under which people form beliefs about political actors and policies" (PC-32658).

Synthetic representations of politicians pose additional challenges when used to fabricate and misrepresent their positions on contentious issues, such as immigration, as their persuasive effects can be even greater. A high-profile case illustrating this point involved a viral synthetic audio clip of the Mayor of London, the UK capital, Sadiq Khan, misrepresenting his views on Armistice Day commemorations and pro-Palestine demonstrations. The Alan Turing Institute, the UK's national institute for data science and artificial intelligence, reports witnessing AI tools being "increasingly exploited to spread harmful or negative narratives about immigrants in the UK" (see PC-32552 and "Adding Fuel to Fire: AI Information Threats and Crisis Events," February 2026), reflecting global concerns about the intersection of disinformation, hate and its impacts on free expression (see Joint Declaration on AI, Freedom of Expression and Media Freedom, Mandate Holders, October 2025). When disinformation and hatred intersect, the risks of incitement to hostility, discrimination and violence also increase (see hate speech analysis, below).

The majority of the Board also notes that globally, the impacts of this kind of disinformation disproportionately impact already marginalized people, adding to arguments for a more robust social media response to deepfakes generally. Women and gender non-conforming people in the public eye are disproportionately targeted by such fakery (see Tipping Point: Online Violence Impacts, Manifestations and Redress in

the AI Age, UN Women, April 2026). The report found that deepfakes are often deliberate and coordinated, with targeted attacks producing severe mental health consequences and a chilling effect on public participation, with nearly half of women journalists reporting self-censorship in response to such attacks. UN Women, the lead United Nations (UN) entity on gender equality, also documented that deepfakes and gendered disinformation against women are designed to "drive them to deplatform or leave public life altogether." Another UN study reported that image and video-based abuse and the use of AI to create defamatory “synthetic histories” are among the most common vectors of online attacks against women (“Your opinion doesn’t matter, anyway”: Exposing technology-facilitated gender-based violence in an era of generative AI (2nd ed.), UNESCO, April 2023). The Board has previously expressed concerns about the inadequacy of social media companies’ responses to targeted abuse against women in public life, noting that “online attacks have a chilling impact on the free expression of women targeted, diminishing access to information for all, as society hears less from women as a result” (see Account Ban for Targeting Public Figures decision). To be clear: this does not imply creating new or different rules depending on the identity of the target but rather understanding how more robust responses to deceptive AI use generally is important to address these specific and disparate impacts.

For a majority of the Board, a much more robust approach to deceptive use of generative-AI to realistically fabricate the words or actions of politicians on matters of public importance would have three elements: a) increased ability to detect AI-generated content, including through human review; b) clearer public criteria for imposing “high risk” labels; and c) clearer labeling or interstitials that adds friction sufficient to clarify the potential for deception, combined with escalating penalties for users that break the rules, including through content demotion. Such an approach would pose significant benefits to rights of public participation that exceed any burden on the freedom of expression of persons sharing those posts.

a.) Increased ability to detect AI-generated content, including through human review

The majority notes Meta's ongoing investments in automated content moderation have the potential to improve the detection of AI-generated content at scale, including through industry standard provenance signals, as well as improving classifier-based detection. However, user and third-party reporting, combined with human review, should also continue to play a significant role in the detection of harmful AI content. This is especially important for the use of AI intended to deceive the public on important matters. The majority acknowledge that the means for detecting AI use at scale do not solve for the problem of detecting deceptive intent at scale, which will often require contextual assessment by specialists, supported by robust third-party fact-checking.

b.) Clearer public criteria for imposing "high risk" labels

As outlined above, Meta's non-public criteria for imposing "high risk" AI labels is too high. The Board is concerned that this minimalist approach seriously underestimates the implications of intentionally deceptive AI-enabled manipulated media on public discourse. These interventions, which fall far short of removing content, in particular, should not be limited to imminent crises or elections. On the latter, the UN Special Rapporteur on freedom of expression has noted, "the life cycle of electoral disinformation is not limited to the polling but begins long before and persists long after the elections, tainting political discourse and polarizing democratic societies" (A/HRC/59/50, para. 8). The Special Rapporteur has also emphasized that "election integrity and information integrity are closely connected" and that digital platforms have amplified information manipulation "into a tsunami of disinformation, misinformation and hate speech," with consequences that "have been dire" for "political opponents, minorities, migrants and other marginalized groups." (ibid, paras. 3-4). Meta should extend the labeling of misinformation and disinformation beyond just exceptional events. Given that the company is not removing such content, the threshold for imposing labels should be lower.

c.) Increasing friction and penalties

Lastly, the majority of the Board believes the design of Meta's current "high risk" AI labels and related enforcement measures are insufficient to deter the sharing of deceptive AI-generated content that poses serious risks to public participation. This requires Meta to explore how labels, or stronger measures like interstitials, could be used to add friction to users' interaction with deceptive content at the point of viewing it, and not only at the point of sharing it. This may require the wording of labels, which is currently neutral (labels do not directly mention risk of deception), to change to ensure that the purpose is well understood. Users who post such content should be informed of what Meta's rules are, and face consequences beyond the labeling of their posts, which would more effectively deter rule breaking. As with content rated "false" by Meta's third-party fact-checkers, users who post content that is subsequently labelled "High Risk" should be informed the post has been removed from recommendations or demoted in user feeds. Users who post such content should also receive strikes that would allow escalating penalties for repeat offending. This could include feature limits and removal from monetization programs.

A minority of the Board does not view this post as violating Meta's manipulated media rules and is concerned that the majority's position is not sufficiently limited to prevent significant restrictions on protected political speech. International human rights standards recognize the particularly high value of expression in public debate "concerning figures in the public and political domain" (General Comment 34, para. 34). For this minority, consistent with the Board's prior decision in the the Altered Video of President Biden case, the content should receive a an "AI Info" label but not be subject to the "High Risk" label and the additional speech-restrictive features that the majority is recommending Meta to adopt.

Free speech principles have long recognized that in cases like this that involve political speech, speakers must have broad latitude for overstatement and even for falsehood. The logical conclusion of the majority position is that the use of AI to fabricate the views or actions of political figures is so inherently dangerous to democracy and political participation that all such posts must be suppressed through more intrusive measures than Meta currently makes available in its policies, unless the manipulation is obvious to the ordinary user. The precedent of the Altered Video of President Biden case, where the Board found that measures short of removal, such as labeling, may help inform users about content authenticity, cannot be distinguished on the grounds that the fake there was obvious, because the basis for that decision was the availability of less restrictive means (i.e. labeling) to address the harm.

For this minority, it is especially problematic to apply demotion measures or other penalties to speech which follows a familiar form of political satire, as seen in this case. The video here exaggerates a politician's views, taking them to an untrue extreme, for the purpose of exposing their potential for radical consequences. The implicit claim of the video in this case is that the local councillor is so extreme in her desire to "welcome refugees" that she would even welcome those who are rapists. This satire has bite because of highly publicized recent instances where some UK politicians have recently been accused of overlooking instances of violence perpetrated by asylum seekers, including sexual violence against minors. This follows the establishment of a Statutory Independent Inquiry into Grooming Gangs, following a national audit, exploring whether sensitivities around race and ethnicity contributed to failures to protect victims and survivors of sexual offences. To insist that Meta suppress a post because it misrepresents this politician's actual views is to deny protection to an important genre of political speech, on an important issue of intense, current public concern. The video should be labeled as AI-generated, as Board precedent demands, but additional measures to suppress its reach (such as demotion or more aggressive labels) would constitute an impermissible restriction on freedom of expression. This minority also notes that because the "High Risk AI" label is only applied when escalated to Meta's

subject matter experts, and therefore is necessarily only used in an exceedingly small number of cases, use of the more general "AI Info" label would have the virtue of being applied at scale and more equally across different users, viewpoints and contexts. This minority also notes that at the time the video was posted, Meta could not automatically detect AI-use without provenance markers or a flag from Trusted Partners, which this content lacked. This lack of confidence in assessing AI content further underscores the challenges and risks in adopting a more aggressive approach to suppressing content based in part on it being AI-generated.

For another minority, given the gravity of the concerns above, removal is the least restrictive means available to Meta *that are effective*. The Board has repeatedly pointed out Meta's failings at deploying less intrusive "informative" labels over numerous cases (see, for example, AI Generated Video in Iran-Israel Conflict). While the company claims to take these risks seriously, Meta has used these labels for deceptive AI use on an alarmingly small number of posts and appears to seldom (if at all) sanction users for failing to disclose AI use. Measures that exist on paper but are barely deployed are – as a practical matter - not available and are therefore not effective.

At the roundtable the Board held, participants raised significant questions about the effectiveness of labelling to counter the persuasive effects of realistic deepfakes of politicians and their harms on public participation. This is especially concerning when labels are not coupled with limitations on reach, nor penalties on users who repeatedly engage in this behaviour to both educate and deter recidivism. These concerns are greater when other less intrusive measures, such as third-party fact-checking, are also not operating with the support or scale required to address the challenge. For this minority, the concern is not about technical detection capabilities, but rather Meta's policy choice to have such a high threshold for labeling that the so-called "high-risk" labels are essentially never applied, never mind its decision not to couple this labeling with other limitations on reach, or bolster other penalties for deceptive AI-use at scale. Given that faked documentation of politicians' supposed words or actions has little to no public interest value, these members view removal as proportionate: the burden it

imposes on the speaker is significantly less than the benefit of removal to public participation and safeguarding democracy.

Notwithstanding these differences of opinion, the Board also finds that as a matter of legality, Meta's rules on misinformation and the application of labels could be made clearer in three respects (General Comment No. 34, para. 25; report A/HRC/38/35, para. 46). First, there is a disconnect between the broad assertion in the policy rationale that Meta removes "content that is likely to directly contribute to interference with the functioning of political processes" and the rules that follow, which prohibit only a narrow set of "voter or census interference" claims, some of which are U.S. specific. The Misinformation policy rationale should make clearer that it only removes direct misinformation about the means and safety of participating in elections, and further revise country-specific examples from its rules to be globally relevant.

Second, there is a concern that the policy does not specify the penalties for failure to disclose AI use, or the conditions under which those penalties are imposed, requiring further details. Third, the rules on manipulated media provide no detail on how Meta defines a "particularly high risk of materially deceiving the public on a matter of public importance." While a majority of Board Members support more robust measures, and a lower threshold for imposing them (as reflected in recommendations), the Board agrees that greater clarity here, supported by more transparency disclosures on the frequency these measures are deployed, would be very useful to public understanding of Meta's rules and its enforcement efforts.

Hate Speech

The Board finds that the video violates the Hateful Conduct Community Standard. For a majority of the Board, it alleges serious criminality and predatory sexual behavior against refugees as a group based on who they are without qualification. The target is expressly named ("*refugees* are welcome here") and serious criminality is alleged

against them as an entire group ("even if *they* rape our women"). The statement meets the definition of a generalization and builds upon decades of demonization of immigrants as a safety threat in the UK, especially to women and girls. This finding is consistent with the Board's Criticism of EU Migration Policies and Immigrants decision. In that case, a majority found that an AI-generated image of immigrants with text stating that Germany does not need any more "gang rape specialists" was violating on similar grounds.

For the majority, removal of the post under the Hateful Conduct rules is consistent with Meta's human rights responsibilities. The Community Standard is clear, complying with the principle of legality, and removal is necessary and proportionate to protect the rights of others to equality and non-discrimination (Article 19, para. 3, ICCPR; Article 2, para. 1, ICCPR).

Removal is proportionate because the benefit of protecting refugees from discrimination, hostility and violence outweighs the burden on the user's freedom of expression (General Comment No. 34, para. 34). Under the UN Rabat Plan of Action and its six factors for assessing the proportionality of restrictions on incitement to violence, hostility or discrimination, these risks are real and near-term.

The social and political context is especially important in this case. In the UK, there has been frequent violence targeting perceived refugees or migrants, as summarized by the UN Committee on the Elimination of Racial Discrimination (Concluding Observations on the UK, UN Committee on the Elimination of Racial Discrimination, 2024), and documented in the Posts Supporting UK Riots case. As that case demonstrated, violence has been fueled by online disinformation, which civil society organizations have continued to document since the Board published that decision (World Report 2026, Human Rights Watch, February 2026; and State of Hate report, Hope Not Hate, March 2026).

Meta's distinctions between inherent traits and behavioral statements in its enforcement guidance is overly mechanical: "refugees rape" is permitted, "refugees are rapists" is not. It turns more on grammar than intent, content or likelihood of harm, which are all relevant factors in the Rabat Plan of Action. The intent behind this post is to demonize refugees as a group, and to promote beliefs regarding the propensities of refugees toward violence. For the majority, the reference to white people also engaging in rape is flippant and aimed at mocking those showing solidarity with refugees; in a context where white people are not in actuality facing this harmful stereotype, it does not have parallel discriminatory impacts.

The cumulative effect of leaving content like this on Meta's platforms creates an environment where violence and discrimination against refugees and ethnic or religious minorities can easily be sparked following triggering events (see Posts Supporting UK Riots). For the majority, Meta's human rights responsibilities are not limited to only preventing widespread racist violence, but also day-to-day acts of hostility and discrimination that precede such events. As the majority in the Criticism of EU Migration Policies and Immigration said, "both the challenge of assessing the impact of each piece of content at scale and the unpredictable nature of online virality justify Meta taking a more cautious approach to moderation ... Meta allowing all hate speech that falls short of incitement as foreseen under Article 20 of the ICCPR would make Meta's platforms an intolerable and unsafe place for minorities and marginalized groups to express themselves." For the majority, removal is necessary because there is no less restrictive measure which would achieve a sufficiently protective function.

For a minority of the Board, the statement in the video does not violate the Hateful Conduct standard, for three separate reasons. First, the video is a (manipulated) depiction of something a politician is imagined to have said, rather than being a direct assertion about refugees. The message of the video is that this politician does not care whether refugees are rapists or not; she would welcome them all. That is a message about the politician, not about refugees as a class.

Second, even taking the words attributed to the politician as a direct assertion, the video does not say that all refugees, or even most refugees, or even large numbers of refugees, are rapists. The claim in the video is that this politician welcomes all refugees without regard to whether they are rapists or not, because “white people” commit rapes too. The post only makes sense as an assertion that the politician is wrong not to distinguish between refugees who are dangerous and those who are not, which would be especially obvious to users in Britain given extensive public debate around these issues (see above).

Third, the standard for removing political speech on the basis of the Hateful Conduct policy is, and ought to be, high. Especially in the context of policy debates that will often touch upon generalizations about groups of people most affected by those policies, which may appear offensive and discriminatory. In the minority’s opinion, Meta’s distinction between prohibited generalizations about the inherent qualities of certain groups and behavioral statements about the conduct of some members of those groups is consistent with international principles of freedom of expression. The majority position, which depends on the concept of the “cumulative harms” of hate speech, largely abandons the need under international human rights law to show direct causation of harm to justify restrictions on speech, and that the harm be likely and imminent (see the minority critique in Criticism of EU Migration Policies and Immigration). The reliance on a notion of cumulative harms without clearly showing the likelihood and imminence of harm is thereby similar to restricting expression simply because it conveys an unfavorable viewpoint, which is manifestly contrary to international human rights norms.

Bullying and Harassment

The Board finds that the video of the politician does not violate the Bullying and Harassment Community Standard because the depicted politician is a public figure and does not benefit from the prohibition on "unwanted manipulated imagery."

Under international human rights standards, political leaders and public officials are, rightly, required to tolerate a higher degree of scrutiny and criticism than private individuals, given their influential role in public affairs and the public's right to hold them accountable (see General Comment No. 34, para. 11, 38). Distinguishing public figures from private individuals helps ensure that any resulting restrictions on expression abide by the principles of necessity and proportionality.

At the same time, the Board notes that Meta's definition of "public figures" in the Community Standard requires clarification. It states that public figures are "state and national level government officials, and political candidates for those offices," and does not include all public elected officials, including at local or city levels. In the Board's view, it is more appropriate for these individuals to be considered public figures due to their elected public role, rather than based on their social media reach or the degree of media reporting about them. Meta appears to be taking that position in its enforcement (as its submissions in this case reflect), but its public rules should communicate this more clearly.

6. The Oversight Board's Decision

The Board overturns Meta's decision to leave up the content, requiring it to be removed under the Hateful Conduct Community Standard. In considering this case, the Board also finds shortcomings in labeling policies under the Misinformation Community Standard and makes recommendations for improvements below.

7. Recommendations

A. Content Policy:

1. To protect users' rights to public participation and information integrity, Meta should change its manipulated media policy for "high risk" labels to lower the threshold for applying these labels and make sure it is not a crisis or election integrity measure only.

   The Board will consider this implemented when Meta provides the updated enforcement guidance to the Board showing the lowered thresholds for applying the "high risk" labels.

2. To protect users' rights to public participation and information integrity, Meta should increase friction for users viewing content labeled as AI-generated or "High Risk AI," such as through an interstitial instead of a label.

   The Board will consider this implemented when Meta provides evidence that increased friction, such as through interstitials, has been implemented for this use case.

3. To protect users' rights to public participation and information integrity, all "High Risk" labeled content should be subject to demonetization, demotion or removal from recommendations, with escalating penalties for accounts that repeatedly share "High Risk" labeled content.

   The Board will consider this recommendation implemented when Meta updates its manipulated media policy to reflect these changes and provides evidence to the Board that such penalties are being applied.

4. To clarify that its Misinformation policy applies to users globally, Meta should ensure that the voters and census interference rules and examples are phrased as globally applicable and not limited to the U.S. context.

The Board will consider this implemented when Meta updates its Misinformation Community Standard to ensure all examples are globally applicable.

B. Transparency:

5. To improve transparency regarding the enforcement of its manipulated media policy, Meta should publicly disclose annual data in its Transparency Center on the number of times it applies the "High Risk" and "High Risk AI" labels to content, including the number of times these labels are applied to content from or about politicians.

   The Board will consider this recommendation implemented when Meta publishes this information in its Transparency Center in a similar format to its annual disclosures on the newsworthiness allowance.

6. To support research into the effectiveness of misinformation interventions, Meta should ensure that all labels under the Misinformation policy ("AI Info," "High Risk" or "High Risk AI") can be seen on content in the Meta Content Library, and that content that has received these labels is searchable via a filter.

   The Board will consider this recommendation implemented when Meta demonstrates that this functionality is operational.

The Board also reiterates the importance of its previous recommendation (no.3) in the AI Generated Video in Israel-Iran Conflict decision, calling for Meta to 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.

***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, the Board did not commission independent research.