AI-Generated Video of Female Campaigner

2026-043-FB-UA

Summary

The Oversight Board is requiring that Meta remove from Facebook an AI-generated video that bullies a young Muslim woman by manipulating her likeness and was part of a viral series of similar posts across social media. The Board finds that Meta's bullying and harassment policies are inadequate to address AI-manipulated imagery of people not normally in the public eye.

Why This Matters

This decision comes as generative AI creates new risks for women and girls, with abuse increasing in scale, speed and severity. According to a 2025 report by UN Women, the lead United Nations (UN) entity on gender equality, many attacks are "often deliberate and coordinated and aimed to silence women in public life." AI-generated videos are changing the nature of online harassment and have significant impacts on the rights of Meta's users, including the public participation rights of Muslim women, and women and girls generally. These attacks often falsify not just an individual's appearance but also their actions and words in a manner that is damaging to them.

About the Case

In 2025, a young Muslim woman in Europe was interviewed on TV as a campaign volunteer for improving menstrual health education and reducing stigma around health conditions for women and girls from ethnic minority backgrounds. In the following months, accounts across social media platforms digitally manipulated the interview footage to ridicule her by creating an apparent disparity between her role

promoting health and her appearance, in an obvious attempt to suggest that she is an unfit messenger on health matters. These videos use her likeness to manipulate her actions and words. The AI-generated videos and images together garnered tens of millions of views across platforms, including Meta's.

The Board selected a Facebook post, submitted by user appeal, as just one example of the abuse that unfolded for this woman. The post includes an AI-generated video of the woman introducing herself as her country's new health ambassador. This false claim is replicated in the caption, alongside a laughter emoji. Her likeness is shown saying that she is going to demonstrate how to stay in shape and then is depicted engaging in a range of limited or absurd exercises and giving advice about healthy eating while consuming junk food. Multiple comments beneath the video from various users ridicule the woman, mainly for her weight and for wearing a hijab.

For the user who appealed to the Board, both their initial report to Meta and subsequent appeal to the company were automatically closed without being prioritized for review. Meta has since confirmed to the Board that its decision to ignore the post, leaving it on the platform without attaching an AI label, was consistent with its policies. Only when the Board asked questions about four disparaging comments on the post did Meta take action to remove the comments, confirming they violated the Bullying and Harassment prohibition on dehumanizing comparisons.

Key Findings

A majority of the Board finds that the video should be removed because it violates Meta's Bullying and Harassment policy. Specifically, it contains a statement of inferiority about a private individual's physical appearance, used AI to falsify her words and actions without her consent, and was part of a campaign that was experienced by the target as mass harassment.

For the majority, Meta's definition of "unwanted manipulated imagery," which is one of the bullying and harassment prohibitions, is too narrow, and should include falsifying a person's actions and words, in addition to altering a person's appearance.

Meta also defines mass harassment campaigns too narrowly, according to the majority, and the company should adopt new measures to reduce the burden on victims reporting them. These measures include adding violating unwanted manipulated content to the company's Media Matching Bank so identical and near-identical content is removed, and using signals of AI-generated or manipulated media as an indicator of severity to prioritize reports for review.

A minority of the Board found the content did not meet Meta's definition of Bullying and Harassment, and that leaving the content up conformed with the company's human rights responsibilities.

The Oversight Board's Decision

The Board overturns Meta's decision to leave up the content, requiring the post to be removed.

The Board stresses the importance of its previous recommendations calling for allowing third-party accounts to report bullying and harassment content.

Additionally, it recommends that Meta:

- Broaden the public definition of "unwanted manipulated imagery" to include deepfakes of a private individual saying or doing things they did not say or do.
- Add instances of violating "unwanted manipulated imagery" to Media Matching Banks for removal.
- Use signals of AI-generated or manipulated media as an indicator of severity for the purpose of prioritizing reports for review.

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

Full Case Decision

1. Case Description and Background

In 2025, a young Muslim woman was interviewed on the news in Europe as a volunteer for a campaign speaking about improving menstrual health education and reducing stigma around health conditions for women and girls from ethnic minority backgrounds. In the following months, this footage of the young woman was digitally manipulated by accounts across various social media platforms to ridicule her by creating an apparent disparity between a role promoting health and her appearance. The imagery was an obvious attempt to spotlight her weight and suggest that she is an unfit messenger on health matters. The AI-generated videos and images included a range of dehumanizing and degrading scenes, including montages of her eating large amounts of food, struggling to work out at a gym and depicting her as an animal, that together garnered tens of millions of views across platforms, including Meta's.

The post in this case is just one example of the abuse that unfolded targeting the woman. It includes an AI-generated video where the woman is depicted out of context introducing herself as her country's new health ambassador. This false claim, which fact-checking organizations debunked in relation to other posts, is replicated in the caption, alongside a laughter emoji. In the AI-generated video, her likeness is shown saying that she is going to demonstrate how to stay in shape and then is depicted engaging in a range of limited or absurd exercises and giving advice about healthy eating while consuming junk food. In the original interview, she did not speak about healthy eating or exercise but addressed the importance of menstrual health education around conditions like endometriosis for women and girls.

The post has been viewed over 5,000 times, received over 200 reactions, and over 50 comments. Multiple comments beneath the video from various users ridicule the woman's appearance, mainly for her weight and for her wearing a hijab.

For the user who appealed to the Board, their reports were not prioritized for review. Meta has since confirmed to the Board that leaving the post on the platform without attaching an AI label was consistent with its policies. Only when the Board asked questions about four disparaging comments on the post did Meta take action to remove them, confirming they violated the Bullying and Harassment prohibition on dehumanizing comparisons. In response to Board questions, Meta said it did not identify any other instances in which it actioned deepfake videos of the woman depicted during the relevant period.

Generative AI has created new risks online, especially for women and girls, with abuse increasing in scale, speed and severity. A UN Women report published in 2025 found that 38% of women in 51 countries have personal experiences of online abuse, which can include cyber-harassment. The report emphasizes that "mainstream AI tools can ... have unintended consequences in amplifying bias or intensifying violence and abuse. For instance, mainstream AI tools can be used to generate disinformation campaigns or to create and disseminate hateful content and harassment campaigns automatically and at scale." Many attacks are "often deliberate and coordinated and aimed to silence women in public life."

A 2026 report from the UN Working Group on Discrimination Against Women and Girls explained: "technology-facilitated gender-based violence is not an isolated set of online abuses but a structural tool of discrimination that restricts women's and girls' participation in public, political and civic life. Virtual violence often mirrors and amplifies patterns of violence and repression in physical public spaces. By driving women and girls out of digital spaces, technology-facilitated gender-based violence

entrenches inequality, undermines autonomy and weakens the conditions necessary for inclusive democratic engagement, threatening their human rights in contradiction with the requirements of Article 19 of the International Covenant on Civil and Political Rights...” (A/HRC/ 62/48, para. 49). The UN’s human rights office, OHCHR, reported in 2026 that 70% of women in North America and Europe have encountered instances of technology-facilitated gender-based violence, which includes online harassment. Public comments in this case also highlight the specific challenges faced by women who speak out on topics that are considered “taboo” in some contexts related to gender and sexual health (see PC – 32637 – Charlotte Manson, City Law School).

The UK Parliament's Women and Equalities Committee documented in January 2026 that Muslim women visibly identifiable by their hijab are disproportionately targeted for harassment both online and in public, and they face a compounding "triple penalty" of race, gender and faith-based harassment. This leads to “self-censoring behaviors and a withdrawal from participation in public life,” the committee said.

2. User Submissions

In their appeal to the Board, the user who reported the content asking for its removal said it used AI to generate a woman’s likeness without her consent, to fabricate misleading content and publicly shame her for her advocacy.

3. Meta’s Content Policies and Submissions

Meta’s Bullying and Harassment Community Standard states that bullying and harassment can come in many forms, and that the company does “not tolerate this kind of behavior because it prevents people from feeling safe and respected...”. The policy provides four tiers of protection based on the severity of the attack, the public figure status of the target and whether they are an adult or minor. While all people are

protected against the most severe attacks (Tier 1), private individuals, limited scope public figures, and minors receive additional protections against less severe attacks (Tiers 2 – 4). In other words, Meta's policy calibrates protection for those who are less in the public eye and/or vulnerable due to their age.

Statements of Inferiority About Physical Appearance

Everyone is protected from "statements of inferiority about physical appearance" under Tier 1 of the Bullying and Harassment policy. Tier 1 is the most severe category of abuse that protects all people – public figures, private individuals and minors. Meta found the post did not violate this rule. While the juxtaposition of health-related activities with the consumption of unhealthy food could be read as implicitly commenting on how the depicted woman looks, Meta found that the post did not contain an explicit statement attacking or directly referencing her physical appearance as inferior.

Unwanted Manipulated Imagery

"Unwanted manipulated imagery" is removed under Tier 3 when three conditions are all met.

First, this protection is only granted to private individuals or minors who are involuntary public figures. Meta's policy rationale states this level of protection goes further to protect private individuals from being "degraded" or "shamed." Meta justifies this public figure distinction because it wants "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 defined in the policy as "state- and national-level government officials, political candidates for those offices, people with over 1 million fans or followers on social media and people who receive substantial news coverage." So-called "limited scope public figures," which include "individuals whose primary

fame is limited to their activism or journalism, or those who become famous through involuntary means," are also not protected from Tier 3 attacks. In this case, Meta says that it treated the targeted woman as a private individual, and not as a public figure or limited scope public figure, so she was eligible for protection from unwanted manipulated imagery.

Second, the target must indicate that the manipulated imagery is unwanted by self-reporting the content. The policy states that self-reporting helps the company "understand that the person targeted feels bullied or harassed." Meta confirmed that the targeted woman did not report this post, so this condition was not met.

Third, the content must be "manipulated imagery," which is not publicly defined in the policy. However, Meta informed the Board that this rule only addresses digital manipulations that alter a person's appearance, not depictions of their actions. In this case, the content features synthetic depictions of the woman saying things she did not say and doing things she did not do. According to Meta, there were no digital alterations of her appearance, so it did not constitute "manipulated imagery" under the policy.

Targeted Mass Harassment

Outside of the four tiers of protection, Meta's Bullying and Harassment Community Standard also states that it may remove "directed mass harassment" when it targets "individuals at heightened risk of offline harm," including human rights defenders. Meta states this policy requires "additional context to enforce" (i.e., the policy is not enforced at scale, but only by specialized internal teams). While Meta notes that the targeted individual meets their definition of a human rights defender, they did not find the post and similar posts against her to qualify as directing others to engage in harassment. According to Meta, there were no organized efforts to direct people towards her profile, or to direct people to collectively produce violating content.

The Board asked Meta questions on reporting mechanisms for bullying and harassment content, coordinated online harassment campaigns, and enforcement of the rules for AI-generated bullying and harassment videos. Meta responded to all questions.

4. Public Comments

The Board received three public comments that met the terms for submission. One of the comments was submitted from Central and South Asia, one from the United States and Canada, and one from Europe. To read public comments submitted with consent to publish, click here.

The submissions covered, among others, the following themes: gendered disinformation and AI-generated harassment, Meta's reporting requirement for bullying and harassment content, the impact of online harassment on public discussions and human rights advocacy, and barriers to sexual health information.

5. Oversight Board Analysis

The Board selected this case to examine how AI-generated videos are changing the nature of online harassment, their impacts on the rights of Meta's users and the implications for the expression and participation rights of women and girls. The case is relevant to two of the Board's seven strategic priorities, Automation and AI and Gender.

5.1 Compliance With Meta's Content Policies

I. Content Rules

For the majority of the Board, this video violates the Bullying and Harassment Community Standard on several fronts. In the policy rationale for its Bullying and Harassment Community Standard, Meta states it removes content that is "meant to

degrade or shame" private individuals. It is challenging to interpret the purpose and effect of this content as anything other than intentionally degrading and shaming a woman based on her appearance and in reaction to her advocacy on women's health issues.

Statements of Inferiority About Physical Appearance

The majority finds that the AI-generated video constitutes a violating "statement of inferiority" about the targeted woman's physical appearance. It falsely proffers that she holds a government office for health (an "ambassador" for her country) to stage an apparent contradiction between this claim and her body size and supposed physical health. By realistically depicting her likeness engaged in absurd exercise routines and eating junk food, the video makes a very clear and explicit claim of inferiority based on her appearance aimed at humiliating her.

For the majority, Meta's failure to interpret this content as an "explicit" statement of inferiority is troubling, in part because its own rules do not specify "explicitness" as a requirement for finding a violation. Though finding a violation in this case relies on an inference drawn from contrasting a false statement with a visual representation, the effect is neither ambiguous nor subtle. The increased accessibility and capabilities of generative AI tools turbo-charge the ability to manufacture and share baseless allegations of hypocrisy. The Board has previously questioned Meta's poor application of its rules to visual content, being too literal or weak on contextual analysis when not finding direct violations in video or images (see Knin Cartoon, Post in Polish Targeting Trans People, Content Targeting Human Rights Defender in Peru). The disparity between Meta's enforcement of visual content and written content is demonstrated here by Meta's removal, following Board questions, of four comments beneath the post that made written dehumanizing comparisons targeting the woman. Those comments, in the Board's view, were no more explicit than the message of the video, creating an appearance of arbitrariness in Meta's enforcement. Given that Meta's platforms are

increasingly designed around promoting visual content, and that moderation shapes how people communicate on its platforms, this approach of only removing "explicit" statements of inferiority is flawed. Veiled attacks, threats and insults are widespread on social media, and Meta's inability or refusal to grapple with them creates gaping holes in enforcement and protection for users from bullying and harassment.

While some might claim the post is an attempt at satire or humor, Meta has chosen not to create an exception for humor in the Bullying and Harassment Community Standard. Any hypocrisy this post attempts to highlight between her fabricated public role (she was not, in fact, advocating healthy eating or exercise) and physical appearance is clearly done for the purpose of bullying and ridiculing a person because of how she looks.

Unwanted Manipulated Imagery

For the majority of the Board, if the depicted woman had discovered and reported the post, it would have violated the unwanted manipulated imagery policy. The rule treats self-reports as the only indicator of manipulated imagery being "unwanted" or non-consensual. This puts the onerous burden of self-reporting on targets of bullying (see PC-32633 – Center for Reproductive Rights; PC-32637 – Charlotte Manson, City Law School), and this condition was not met. Nevertheless, the majority takes issue with Meta's interpretation that even if the woman had reported the post, it would still not violate this rule. A plain reading of "manipulated imagery" encompasses fabrication of a person's words or actions and is not limited to manipulation of a person's physical appearance. It is unclear what purpose Meta's distinction between manipulation of appearance and action serves. Fabrication of words and conduct, as seen in this case, can degrade or shame individuals for their appearance just as severely (if not more) as fictitious and exaggerated representations of their appearance (see PC-32635 – Digital Rights Foundation). As seen in the Altered Video of President Biden case on Meta's misinformation policy, enforcement guidance reacting to the content trends of one era

can quickly be eclipsed by technological advancement. Developments in hyper-realistic generative AI video, even since the Board decided the Biden case, indicate Meta's rules on Bullying and Harassment need updating too.

Targeted Mass Harassment

The Board agrees with Meta that the depicted campaigner is a human rights defender and acknowledges this post did not explicitly coordinate or give instructions to direct harassment against her. However, the majority of the Board finds that Meta's conclusion rests on an overly narrow reading of what "directed" mass harassment entails in the digital age. Several prominent social media accounts engaged in the abuse, essentially directing adverse attention towards her, and effectively leading to a pile-on that the post in this case was part of, with all the hallmarks of mass harassment. From its own research, the Board observed that content sharing the false claim that this woman held an official government position peaked at around 250 in a single day, totaling more than 1,000 over the period studied. The similarity of many clusters of posts sharing identical or near-identical captions provides evidence of coordination and would certainly have been experienced as mass harassment by the woman targeted. This level of coordination is further evidence of harassing intent, and points to a structural problem with Meta's enforcement systems not providing users with specific tools to collectively report multiple posts as part of a violating campaign. The Board also noted examples on the social media platform X where users asking the AI chatbot Grok if the false claims were real received responses that the false claims were true.

The Board agrees that Meta correctly removed the comments beneath this post that engaged in direct dehumanizing comparisons, albeit after the Board brought them to Meta's attention. For the majority, the nature of these comments shows the audience understood the parent post and chose to join its author in mass harassment of the targeted woman. Other members of the Board accept these were violating comments

but note the author of a post should not have their content evaluated based on the comments a post elicits from others.

For a minority of the Board, the post does not violate the Bullying and Harassment Community Standard as it does not contain a "statement of inferiority about physical appearance." To interpret this rule as covering inferences, as the majority does, is excessively broad, asserting one contextual interpretation above other equally plausible ones. The video addresses an ongoing public conversation about the seeming contradiction between a person appearing on the news to advocate healthy living and that person not appearing to live up to that standard. For the minority, this style of calling out perceived hypocrisy may be cruel or unkind, but it did not go as far as stating inferiority, i.e., that she is *less than*, due to her physical appearance. Furthermore, for a minority of the Board, the targeted woman does not benefit from Tier 3 protections for unwanted manipulated imagery, not for the reasons Meta states, but because she is a "limited scope public figure" and not a private individual. This is due to her public-facing role for an advocacy organization and appearance on national news media in this capacity. For these members of the Board, the volume of similar posts is not material for finding a violation for targeted mass harassment, since this post does not reach Meta's threshold for violating the Bullying and Harassment policy.

5.2 Compliance With Meta's Human Rights Responsibilities

The majority of the Board finds Meta's human rights responsibilities support removal of this post. A minority finds that they do not.

1. *Freedom of Expression (Article 19 ICCPR)*

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, International Covenant on Civil and Political Rights (ICCPR)). These

requirements are often referred to as the “three-part test.” The Board uses this framework to interpret Meta’s human rights responsibilities in line with the UN Guiding Principles on Business and Human Rights, which Meta itself has committed to in its Corporate Human Rights Policy. As the UN Special Rapporteur on freedom of expression has stated, although “companies do not have the obligations of governments, their impact is of a sort that requires them to assess the same kind of questions about protecting their users' right to freedom of expression” (A/74/486, para. 41).

I. Legality (Clarity and Accessibility of the Rules)

The principle of legality requires rules limiting expression to be accessible and clear, formulated with sufficient precision to enable an individual to regulate their conduct accordingly (General Comment No. 34, para. 25). Additionally, these rules “may not confer unfettered discretion for the restriction of freedom of expression on those charged with [their] execution” and must “provide sufficient guidance to those charged with their execution to enable them to ascertain what sorts of expression are properly restricted and what sorts are not” (ibid). The UN Special Rapporteur on freedom of expression has stated that, when applied to private actors’ governance of online speech, rules should be clear and specific (A/HRC/38/35, para. 46). People using Meta’s platforms should be able to access and understand the rules and content reviewers should have clear guidance regarding their enforcement.

The majority of the Board found Meta’s rules to be sufficiently clear to reach the outcomes in this case but note three areas where Meta’s confidential internal guidance requires updating to reflect how its rules are framed to the public. This could help ensure more accurate enforcement and reduce perceptions of arbitrariness in Meta’s interpretation of its rules.

First, there should be clear guidance to reviewers that "statements of inferiority about physical appearances" should include violations made through images and video as well as in the written or spoken word. Specific examples for reviewers of how deepfakes can fit within this category would be helpful.

Second, the arbitrariness that flows from the disconnect between the letter of the policy and the enforcement guidance will be compounded as AI-generated content becomes more ubiquitous and implies a policy change is needed. Meta should make clear its policy on unwanted manipulated imagery encompasses non-consensual manipulation of a person's words or conduct, and the public rule should be updated to make this explicit. If a person is falsely depicted in a manipulated image as humiliated, degraded or doing something illegal, the policy should be triggered.

Third, Meta should make the requirements of its rule on "directed mass harassment" clearer. As outlined below, Meta should improve its approach to detecting and removing mass harassment campaigns of the kind this case represents, which may require a combination of policy and enforcement changes.

II. Legitimate Aim

Any restriction on freedom of expression should also pursue one or more of the legitimate aims listed in the ICCPR, which includes protecting rights of others (Article 19, para. 3, ICCPR). The Bullying and Harassment Community Standard seeks to protect the rights of others, including their right to privacy, reputation and public participation (Articles 17 and 25, ICCPR) and to physical and mental health (Article 12, International Covenant on Economic, Social and Cultural Rights).

III. Necessity and Proportionality

Under ICCPR Article 19(3), necessity and proportionality requires that restrictions on expression "must be appropriate to achieve their protective function; they must be the least intrusive instrument amongst those which might achieve their protective function; they must be proportionate to the interest to be protected" (General Comment No. 34, para. 34).

It is important to start this analysis with the rights of the woman who was targeted in this harassment campaign, and the rights of people to receive the important sexual and reproductive health information she appeared on the news to share. It has long been documented that women who engage in public discourse online, especially if they are in the public eye and belong to minority religious or ethnic groups, are exposed to much higher volumes of harassment and abuse on social media than other people (see PC-32635; "#Toxictwitter: violence and abuse against women online," Amnesty International report, November 2018; "No Excuse for Abuse," Pen America report, March 2021).

As realistic generative AI video capabilities rapidly develop to allow people to more easily mimic others online, the application of these tools to bully and harass people who hold unpopular or disfavored views, as well as members of marginalized groups, has increased. This does not necessarily generate novel bullying and harassment harms requiring different norms to resolve, but it has created additional challenges of scale, believability of content created by rapidly improving generative AI apps, and the more acute impacts that come from bullying or harassing a person by using hyper-realistic and deceptive likenesses to inflict harm. These effects are compounded by platform incentives, in particular engagement-based algorithms and monetization programs.

For a majority of Board Members, Meta's human rights responsibilities require a more robust approach, so that content like the AI-generated video in this case would be removed for violating the Bullying and Harassment policy. For a minority, while also concerned by the effects of bullying and harassment, measures less intrusive than

removal, such as labelling, demotion or de-monetization could be utilized if content of this sort was found to be violative. The minority sees these measures as more in line with Meta's human rights responsibilities and necessary to guard against overbroad restrictions on expression.

For the majority of the Board, removal is necessary because there is no less restrictive measure which would achieve a sufficiently protective function. The harm in this case does not flow primarily from deception, but from the weaponization of a person's realistic likeness as a tool for humiliating and misrepresenting them. This cannot be mitigated through warning screens, labeling or the demotion of content as these measures would allow the harassing content to still exist on the platform. Moreover, labels are seldom applied and do not increase friction or slow the dissemination of such content unless rated by third-party fact-checkers (see AI-Generated Content in Israel-Iran Conflict).

Removal is also proportionate because the benefits of the restriction to the targeted woman's privacy and reputation, as well as to her right to public participation, are greater than any burden on the expression of the user whose content is removed.

There is little to no public interest in the speech being unrestricted. The majority of the Board finds that the woman has not engaged in wrongdoing or hypocrisy; the content maliciously misrepresents her position to attack her and does not engage with the substance of her arguments. Moreover, the target of the abuse is a private individual. She is not a public figure holding office, or otherwise in a position of power that demands she show a heightened tolerance for such attacks (General Comment No. 34, *op. cit.*, at para. 38). Even following the removal of this post, the user who posted the content remains free to discuss the same issues the woman went on the news to discuss, without engaging in irrelevant and personalized attacks. Conversely, there is significant public interest in the information the woman shared on the news – on sexual and reproductive health, and the related critique of government education policy.

The Board's research revealed that posts similar to the content in this case, several of which were also AI-generated, appeared across many separate accounts on Facebook and Instagram, as well as on non-Meta platforms, often with identical and near-identical captions. While many other social media platforms have taken a similarly opaque position as Meta in their approach to comparable AI-generated content that harasses an individual (including on X and YouTube), some have been more explicit with their guardrails. For example, TikTok has rules that prohibit AI-created likenesses (defined as "a recognizable image, video or audio representation of a person, including their face, body, voice and gestures") made to bully or harass an individual and require disclosure of AI-generated content that makes it look like someone is saying or doing something that they did not. Such abuse, especially when produced at high volume and across platforms, has a psychological toll on those targeted that can impact their health, be a precursor to in-person harassment and violence, and can effectively force people to withdraw from public discourse. This is further exacerbated for individuals such as the woman depicted, who are speaking out on sexual and reproductive health issues. This chills not only the target's expression – affecting the right to receive information of the individual's intended audience – but also, potentially, the expression of bystanders who witness the abuse and internalize the risks that flow from engaging in public discourse. A permissive environment for such abuse risks not only normalizing the abuse but also advancing the marginalization it seeks to achieve. The harms and impact of such content, including digitally manipulated imagery or videos of a targeted individual, often extend to a whole group and wider community, undermining multiple human rights and weakening inclusive participation in public life through this chilling effect (see PC-32637 – Charlotte Manson, City Law School; A/HRC/62/48). This case builds upon the Reported AI-Generated Sexualized Video decision to make clear that the harms of "deepfakes" are not limited only to harassment in the form of non-consensual sexualized content.

Whereas several Board cases have focused on AI-generated sexualized impersonations of women, this case demonstrates how users' rights can also be significantly infringed through non-sexual synthetic imitation. This deepfake was posted in the context of a broader online campaign of harassment, where the posting user decided to go further than many others in not simply sharing an opinion, but posting a deepfake that is particularly malicious in its degrading and humiliating effects. The aggregate impact of the mass bullying or harassment on this individual's privacy and dignity sought to force her retreat from public discourse.

Cumulatively, such content has societal impacts too, sending a clear message to other women and girls, especially those who look like her, not to speak out. According to public comments the Board received, this case demonstrates "a broader phenomenon in which women and girls who engage publicly on issues relating to health, gender equality and sexual and reproductive rights are disproportionately subjected to gendered harassment, manipulated media and disinformation" (see PC-32633 – Center for Reproductive Rights). This form of AI-generated content can effectively silence women and girls, especially those who are visibly religious, from speaking out on issues that may be considered taboo or controversial in their communities. For the majority, while solidarity with victims through counter-speech should be encouraged, it is not a remedy for bullying and harassment. This is especially the case in parts of the world where government censorship, societal stigma and a lack of resources for civil society or independent media do not guarantee the conditions counter-speech requires. These are not guaranteed conditions for other marginalized individuals on the platform, in addition.

For a minority of the Board, removal was not necessary nor proportionate. While it is reasonable to disagree with the user's behavior in this case and even condemn it, there is still, broadly, a public interest in being free to critically discuss the news and the people who appear on it. That requires defending speech that, as here, is objectionable and even insulting and unkind in some instances. Censorship in this context is only

likely to compound division, especially where people might perceive that criticism of individuals in the public domain is being removed based on a differing viewpoint. The minority notes that the targeted individual in this case has spoken out after this experience, and an impressive number of high-profile people spoke out in her defense, making an example of her abusers through counter-speech. The minority is concerned that were Meta to attempt to scale enforcement of the decision the majority reached in this case, it would lead to overbroad restrictions on freedom of expression, and potential abuse of the Bullying and Harassment policy by people in power to suppress legitimate criticism. Meta's narrow framing of some of its Bullying and Harassment rules is proportionate to prevent that overreach, and international human rights law does not, in the view of the minority, require the company to take a more restrictive approach.

Access to Remedy

As with many prior cases concerning violating use of generative AI videos (see Explicit AI Images of Female Public Figures, Reported AI-Generated Sexualized Video), Meta did not prioritize the content in this case for review, either on first instance or on appeal. The Board is seriously concerned at the volume of user reports against violating content that are assessed by classifiers as warranting human review but receive none. This is especially concerning in cases of bullying and harassment, where failure to assess content severely harms victims, and ignoring valid user reports undermines trust in Meta's content moderation systems. Advancements in the use of automation in content moderation could provide opportunities to significantly reduce the number of user reports that are not reviewed, particularly in more straight forward cases, freeing up reviewer capacity for more contextually complex or specialist reviews.

The Board reiterates its concern that harassment victims often have the burden of reporting content, which is compounded by the lack of third-party reporting tools, especially where harassment goes viral (see Gender Identity Debate Videos). These

challenges are especially acute for women in parts of the world – especially for hijabi or other visibly religious women – where their autonomy is restricted through multiple and intersecting forms of social norms, discrimination and even violence.

As Meta updates its policies and guidance to ensure this content is treated as violating, including when it is in visual form and may not contain explicit or literal statements, it is important that its systems are trained to prioritize such posts for review, and that adequate resources are made available in an effective manner to ensure that reports are not auto-closed due to the lack of review capacity. Where individuals are victims of mass-harassment, there should be tools to assist reporting of connected violations, which should be prioritized and reviewed together for full context.

6. The Oversight Board's Decision

A majority of the Board overturns Meta's decision to leave up the content, requiring the post to be removed.

7. Recommendations

A. Content Policy

1. To address the gap in the Bullying and Harassment policy, Meta should publicly define "unwanted manipulated imagery" as including deepfakes of a private individual saying or doing things they did not say or do, as well as manipulation of a person's appearance.

   The Board will consider this recommendation implemented when the policy has been publicly updated with this definition.

B. Enforcement

The Board reiterates its previous recommendation in the Gender Identity Videos Bundle decision calling for allowing third-party accounts to report bullying and harassment content (recommendation no. 3), noting the relevance of this recommendation to the issues of case. The Board is concerned that although it made this recommendation well over a year ago Meta's response is still pending while it assesses recommendation feasibility. The Board encourages Meta to complete this assessment and implement the recommendation.

2. To reduce the reporting burden on victims of mass harassment using deepfakes, Meta should add instances of violating "unwanted manipulated imagery" to Media Matching Banks for removal.

   The Board will consider this recommendation implemented when Meta provides evidence to the Board that it is able to scale reports against deepfakes to identical content.

3. To ensure access to remedy, Meta should use signals of AI-generated or manipulated media as an indicator of severity for the purpose of prioritizing reports for review.

   The Board will consider this recommendation implemented when Meta provides evidence that AI-generation has been integrated as a signal to the prioritization of the Bullying and Harassment human review queues, and shares metrics on any resulting changes to the volume of auto-closed Bullying and Harassment reports.

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.