Explicit AI Images of Female Public Figures

2024-007-IG-UA, 2024-008-FB-UA

Summary

In two cases of explicit AI images that resemble female public figures from India and the United States, the Oversight Board finds that both posts should have been removed from Meta's platforms. Deepfake intimate images disproportionately affect women and girls – undermining their rights to privacy and protection from mental and physical harm. Restrictions on this content are legitimate to protect individuals from the creation and dissemination of sexual images made without their consent. Given the severity of harms, removing the content is the only effective way to protect the people impacted. Labeling manipulated content is not appropriate in this instance because the harms stem from the sharing and viewing of these images – and not solely from misleading people about their authenticity. The Board's recommendations seek to make Meta's rules on this type of content more intuitive and to make it easier for users to report non-consensual sexualized images.

About the Cases

These two cases involve AI-generated images of nude women, one resembling an Indian public figure, the other an American public figure. In the first case, an Instagram account that shared only AI-generated or manipulated images of Indian women posted a picture of the back of a nude woman with her face visible, as part of a set of images. This set also featured a similar picture of the woman in beachwear, most likely the source material for the explicit AI manipulation. The second case also involves an explicit AI-generated image resembling a female public figure, this time from the United States. In this image, posted to a Facebook group for AI creations, the nude woman is being groped. The famous figure she resembles is named in the caption.

In the first case (Indian public figure), a user reported the content to Meta for pornography but as the report was not reviewed within 48 hours, it was automatically closed. The user then appealed to Meta, but this was also automatically closed. Finally, the user appealed to the Board. As a result of the Board selecting this case, Meta determined that its original decision to leave the content on Instagram was in error and the company removed the post for violating the Bullying and Harassment Community Standard. Later, after the Board began its deliberations, Meta disabled the account that posted the content and added the explicit image to a Media Matching Service (MMS) bank.

In the second case (American public figure), the explicit image had already been added to an MMS bank for violating Meta's Bullying and Harassment policy and so was automatically removed. These banks automatically find and remove images that already have been identified by human reviewers as breaking Meta's rules. The user who posted the AI-generated image appealed but this was automatically closed. The user then appealed to the Board to have their post restored.

Deepfake intimate images comprise synthetic media digitally manipulated to depict real people in a sexualized way. It is becoming easier to create, with fewer pictures required to generate a realistic image. One report points to a 550% increase in online deepfake videos since 2019, the vast majority of which are sexualized depictions of real individuals and target women.

Key Findings

The Board finds that both images violated Meta's rule that prohibits "derogatory sexualized photoshop" under the Bullying and Harassment policy. It is clear the images have been edited to show the faces of real public figures with a different (real or fictional) nude body, while contextual clues, including hashtags and where the content was posted, also indicate they are AI-generated. In the second case (American public figure), there is an additional violation of the Adult Nudity and Sexual Activity policy as the explicit image shows the woman having her breast squeezed. Removing both posts was in line with Meta's human rights responsibilities.

The Board believes that people using Meta's platforms should be able to understand the rules. While the term "derogatory sexualized photoshop" should have been clear enough to the two users posting in these cases, it is not sufficiently clear more generally to users. When the Board asked Meta about the meaning, the company said the term refers to "manipulated images that are sexualized in ways that are likely to be unwanted by the target and thus perceived as derogatory." The Board notes that a different term such as "non-consensual" would be a clearer description to explain the idea of unwanted sexualized manipulations of images. Additionally, the Board finds that "photoshop" is too narrow to cover the array of media manipulation techniques available today, especially generative AI. Meta needs to specify in this rule that the prohibition on this content covers this broader range of editing techniques.

To ensure the rules prohibiting non-consensual sexualized images are more intuitive, the Board finds they should be part of the Adult Sexual Exploitation Community Standard, rather than Bullying and Harassment. In both these cases, users would have been unlikely to perceive them as an issue of Bullying and Harassment. External research shows that users post such content for many reasons besides harassment and trolling, including a desire to build an audience, monetize pages or direct users to other sites, including pornographic ones. Therefore, Meta's rules on these images would be clearer if the focus was on the lack of consent and the harms from such content proliferating – rather than the impact of direct attacks, which is what is implied by enforcing under Bullying and Harassment. The Adult Sexual Exploitation policy would be a more logical place for these rules. This policy already prohibits non-consensual intimate images, which is a similar issue as both are examples of image-based sexual abuse. Then, Meta could also consider renaming the policy to "Non-Consensual Sexual Content."

The Board notes the image resembling an Indian public figure was not added to an MMS bank by Meta until the Board asked why. Meta responded by saying that it relied on media reports to add the image resembling the American public figure to the bank, but there were no such media signals in the first case. This is worrying because many victims of deepfake intimate images are not in the public eye and are forced to either accept the spread of their non-consensual depictions or search for and report every instance. One of the existing signals of lack of consent under the Adult Sexual Exploitation policy is media reports of leaks of non-consensual intimate images. This can be useful when posts involve public figures but is not helpful for private individuals. Therefore, Meta should not be over-reliant on this signal. The Board also suggests that context indicating the nude or sexualized aspects of the content are AI-generated, photoshopped or otherwise manipulated be considered as a signal of non-consent.

Finally, the Board is concerned about the auto-closing of appeals for image-based sexual abuse. Even waiting 48 hours for a review can be harmful given the damage caused. The Board does not yet have sufficient information on Meta's use of auto-closing generally but considers this an issue that could have a significant human rights impact, requiring risk assessment and mitigation.

The Oversight Board's Decision

In the first case (Indian public figure), the Board overturns Meta's original decision to leave up the post. In the second case (American public figure), the Board upholds Meta's decision to take down the post.

The Board recommends that Meta:

- Move the prohibition on "derogatory sexualized photoshop" into the Adult Sexual Exploitation Community Standard.
- Change the word "derogatory" in the prohibition on "derogatory sexualized photoshop" to "non-consensual."
- Replace the word "photoshop" in the prohibition on "derogatory sexualized photoshop" with a more generalized term for manipulated media.

- Harmonize its policies on non-consensual content by adding a new signal for lack of consent in the Adult Sexual Exploitation policy: context that content is AI-generated or manipulated. For content with this specific context, the policy should also specify that it need not be “non-commercial or produced in a private setting” to be violating.

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

Full Case Decision

1. Case Description and Background

The Oversight Board has reviewed two cases together, one posted on Facebook, the other on Instagram, by different users.

The first case involves an AI-manipulated image of a nude woman shown from the back with her face visible. Posted on Instagram, the image resembles a female public figure from India and was part of a set of images featuring a similar picture of the woman, in beachwear, which was likely the source material for the AI manipulation. The account that posted this content describes itself as only sharing AI-generated images of Indian women, and the caption includes hashtags indicating the image was created using AI.

A user reported the content to Meta for pornography. This report was automatically closed because it was not reviewed within 48 hours. The user then appealed Meta’s decision to leave up the content, but this was also automatically closed and so the content remained up. The user then appealed to the Board. As a result of the Board selecting this case, Meta determined that its decision to leave the content up was in error and it removed the post for violating the Community Standard. Later, after the Board selected the case, Meta disabled the account that posted the content and added the content to a Media Matching Service (MMS) bank.

The second case concerns an image posted to a Facebook group for AI creations. It shows an AI-generated image of a nude woman being groped on the breast. The image was created with AI so as to resemble an American public figure, who is named in the caption.

In this second case, the image was removed for violating Meta's Bullying and Harassment policy. A different user had already posted an identical image, which led to it being escalated to Meta's policy or subject matter experts who decided the content violated the Bullying and Harassment policy, specifically for "derogatory sexualized photoshop or drawings," and removed it. That image was then added to an MMS bank. These banks automatically find and remove images that have already been identified as violating. The AI-generated image in the second case was automatically removed because it had been added to an MMS bank. The user who posted the content appealed the removal but the report was automatically closed. They then appealed to the Board to have their content restored.

The Board noted the following context in reaching its decision on these cases.

Deepfake intimate images are synthetic media that have been digitally manipulated to depict real people in a sexualized manner. What is perceived as pornography may differ across countries and cultures. A public comment submitted to the Board from Witness, an international human rights NGO, gives the example of a Bangladeshi deepfake of a female politician in a bikini, which could be particularly harmful because of the cultural context, though it might not be actionable in another cultural setting (see PC-27095).

Deepfake intimate imagery is becoming easier to create using AI tools, with fewer pictures required to generate a realistic image. Women in International

Security (WIIS) explains: “This means that practically everyone who has taken a selfie or posted a picture of themselves online runs the hypothetical risk of having a deepfake created in their image.” *The Guardian* reported that the AI firm Deeptrace analyzed 15,000 deepfake videos it found online in September 2019, noting that 96% were pornographic and 99% of those mapped faces from female celebrities onto porn performers. There has reportedly been a 550% increase in the number of online deepfake videos since 2019, with sexualized depictions of real individuals making up 98% of all deepfake videos online and women comprising 99% of the targeted individuals (Home Security Heroes 2023 report). The top 10 dedicated deepfake intimate imagery websites collectively received more than 34 million monthly visits.

Image-based sexual abuse has been shown to have a significant impact on victims. The UK’s 2019 Adult Online Hate, Harassment and Abuse report quotes a range of studies on image-based sexual abuse (including deepfake intimate imagery) that have examined the experiences of victims. These studies found that victims may struggle with feelings of shame, helplessness, embarrassment, self-blame, anger, guilt, paranoia, isolation, humiliation and powerlessness; along with feeling a loss of integrity, dignity, security, self-esteem, self-respect and self-worth. Researchers of online sexual abuse suggest the harms of deepfake sexual imagery may be as severe as those associated with real non-consensual sexual images.

Deepfake intimate imagery is a global issue. There have been reports of female politicians being targeted in Bangladesh, Pakistan, Italy, , Northern Ireland and Ukraine. Journalists, human rights defenders and celebrities are also routinely targeted. However, anyone can be a victim of deepfake intimate imagery. There have been recent incidents in the United States and Spain of children and young teenagers being targeted with deepfake intimate imagery. Experts consulted by the Board noted that this content can be particularly damaging in socially conservative communities. For instance, an 18-year-old woman was reportedly shot dead by her father and uncle in Pakistan's remote Kohistan region after a digitally altered photograph of her with a man went viral.

Both India and the United States have considered laws and announced further plans to regulate deepfakes. A public comment from Rakesh Maheshwari, a former senior government official in cyber law explains how India's current laws on social media could be applied to the content in the first case (see PC-27029). However, the Board received many public comments emphasizing how important it is that social media platforms be the first line of defense because legal regimes may not move quickly enough to stop this content from proliferating. A public comment from the Indian NGO Breakthrough Trust also explains that in India, "women often face secondary victimisation" when accessing police or court services by being asked why they put pictures of themselves on the internet in the first place – even when the images were deepfaked (see PC-27044).

Meta has been active in developing technologies to address a related issue, non-consensual intimate image sharing (NCII). Independent experts consulted by the Board praised Meta's efforts in finding and removing NCII as industry-leading and called their image-matching technologies valuable. While NCII is different from deepfake intimate imagery in that NCII involves real images whereas deepfakes involve digitally created or altered images, both are examples of image-based sexual abuse.

2. User Submissions

The user who reported the content in the first case said they had seen AI-generated explicit images of celebrities on Instagram and were concerned about this being available on a platform that teenagers were permitted to use. The content creator did not provide a user statement to the Board.

The user who shared the post in the second case stated in their appeal that their intention wasn't to bully, harass or degrade anyone but to entertain people.

3. Meta's Content Policies and Submissions

I. Meta's Content Policies

Bullying and Harassment Community Standard

The Bullying and Harassment Community Standard states under Tier 1 (Universal protections for everyone) that everyone (including public figures) is protected from "Derogatory sexualized photoshop or drawings."

Additional internal guidance provided to content moderators defines "derogatory sexualized photoshop or drawings" as content that has been manipulated or edited to sexualize it in ways that are likely to be unwanted by the target and thus perceived as derogatory – in one example, combining a real person's head with a nude or nearly nude body.

Adult Nudity and Sexual Activity policy

This policy prohibits, among other depictions of nudity and sexual activity, "fully nude close-ups of buttocks" as well as "squeezing female breasts." Squeezing female breasts is "defined as a grabbing motion with curved fingers that shows both marks and clear shape change of the breasts. We allow squeezing in breastfeeding contexts."

Adult Sexual Exploitation policy

This policy prohibits:

- "Sharing, threatening, stating an intent to share, offering or asking for non-consensual intimate imagery that fulfils all of the three following conditions:
    - Imagery is non-commercial or produced in a private setting.
    - Person in the imagery is (near-)nude, engaged in sexual activity or in a sexual pose.
    - Lack of consent to share the imagery is indicated by meeting any of the signals:
        - Vengeful context (such as caption, comments or Page title).
        - Independent sources (such as law enforcement record) including entertainment media (such as leak of images confirmed by media).
        - A visible match between the person depicted in the image and the person who has reported the content to us.
        - The person who reported the content to us shares the same name as the person depicted in the image."

II. Meta's Submissions

Meta assessed both posts under its Bullying and Harassment and Adult Nudity and Sexual Activity policies. The company found both violated the Bullying and Harassment policy, but only the post in the second case (American public figure) violated the Adult Nudity and Sexual Activity Community Standard.

Bullying and Harassment Community Standard

The Bullying and Harassment policy protects both public and private figures from "[d]erogatory sexualized photoshop or drawings," as this type of content "prevents people from feeling safe and respected on Facebook, Instagram and Threads." In response to the Board's question about how the company identifies "photoshopped" or AI-generated content, Meta explained that the assessment is made on a "case-by-case basis" and relies on several signals including "context clues as well as signals from credible sources, such as articles from third party fact-checkers, credible media sources, assessments from onboarded Trusted Partners, and other non-partisan organizations or government partners."

Meta determined that the image in the first case violated this policy because it was created using AI to resemble a public figure from India and was manipulated or edited to make the figure appear nearly nude in a "sexually suggestive pose." The company also considered the user's handle and hashtags, both of which clearly indicate the image is AI generated.

Meta determined that the image in the second case also violated this policy. The company considered the following factors in determining that this image was AI generated: the face appears to be combined with a nearly nude body and the "coloring, texture, and clarity of the image suggested the video [image] was AI-generated"; there was external reporting through the media on the proliferation of such generated images; and the content was posted in a group dedicated to sharing images created using artificial intelligence.

Adult Nudity and Sexual Activity Community Standard

Meta informed the Board that the company determined the image in the second case (American public figure) also violated the Adult Nudity and Sexual Activity policy. Because the content in the second case shows someone "grabbing" the AI-generated image of the female public figure, it violated the prohibition on imagery showing the squeezing of female breasts.

According to the company, the image in the first case (Indian public figure) does not violate the policy because although fully nude close-ups of buttocks are prohibited, this image is not a close-up as defined by the company.

The company also explained that the decision to remove both posts struck the right balance between its values of safety, privacy, dignity and voice because "Meta assessed the creative value of the content in this case bundle as minimal." Looking at the hashtags and captions on both posts, the company concluded the "intent was sexual rather than artistic." The company also concluded that the "safety concern with removing this content outweighed any expressive value of the speech." Citing stakeholder input from an earlier policy forum on "Attacks Against Public Figures," Meta highlighted concerns about abuse and harassment that public figures face online, and argued that this leads to self-censorship and the silencing of those who witness the harassment.

In May 2024, Meta updated its Adult Nudity and Sexual Activity Community Standard, clarifying that the policy applies to all "photorealistic imagery," and that "[w]here it is unclear if an image or video is photorealistic," they "presume that it is." The Board understands this to mean that realistic AI-generated images of real people, celebrities or otherwise, will be removed under the Adult Nudity and Sexual Activity Community Standard when they contain nudity or sexual activity and do not qualify for a narrow range of exceptions. The company cites prevention of "the sharing of non-consensual or underage content" as the rationale for removing photorealistic sexual imagery. The Board welcomes Meta's clarification of the Adult Nudity and Sexual Activity Community Standard, and supports the company's efforts to enforce against realistic-looking fictional images and videos the same way in which it would real ones.

The outcomes for both pieces of content in this case would remain the same under the updated policies: both would still be removed based on the prohibition on derogatory sexualized photoshop under the Bullying and Harassment Community Standard, and the post featuring the American public figure would violate the Adult Nudity and Sexual Activity Community Standard for showing the squeezing of breasts in a context that is not permitted.

While these changes represent a welcome clarification, they are not sufficient to deal with the proliferation of AI-generated non-consensual intimate imagery. The Board reaffirms the importance of a dedicated policy line against AI generated or manipulated non-consensual sexualized content that exists as part of Meta's Adult Sexual Exploitation Community Standard.

According to Meta, its Media Matching Service (MMS) banks identify and act on media, in this case images, posted on its platforms. Once content is identified for banking, it is converted into a string of data or "hash." The hash is then associated with a particular bank. Meta's MMS banks are created to align with specific Community Standard policies, and are not designed around specific behaviors or types of content such as derogatory sexualized photoshopped content. These banks can automatically identify and remove images that have already been identified by human reviewers as violating the company's rules.

The image in the second case (American public figure) was removed because an identical image had already been escalated to human review and added to an MMS bank. The image in the first case (Indian public figure) was initially not added to an MMS bank. Meta stated that: "Not all instances of content found to be violating for derogatory sexualized photoshopping are added to a MMS bank. Requests to bank content must generally be approved, on escalation, by our internal teams. This is because MMS banking is a powerful enforcement tool that may carry over-enforcement risks." Meta only changed its decision to bank the image in the first case after the Board submitted a question asking why this had not been done.

The Board asked 13 questions about MMS Banks, the auto-closing of appeals, the prohibition on derogatory sexualized photoshopping in the Bullying and Harassment policy, and other policies that might be relevant to this case. Meta responded to them all.

4. Public Comments

The Oversight Board received 88 public comments that met the terms for submission. Of these, 14 were from Asia Pacific and Oceania, 24 from Central and

South Asia, 15 from Europe, five from Latin America and Caribbean and 30 from the United States and Canada. To read public comments submitted with consent to publish, click here.

The submissions covered the following themes: the prevalence and cultural implications of deepfakes in India, the impact of deepfake intimate imagery on women generally and female public figures, why auto-closing appeals related to image-based sexual abuse is problematic, and the necessity of combining human and automated systems of review to detect and remove deepfake intimate imagery.

5. Oversight Board Analysis

The Board analyzed Meta's decision in this case against Meta's content policies, values and human rights responsibilities. The Board also assessed the implications of this case for Meta's broader approach to content governance.

5.1 Compliance With Meta's Content Policies

I. Content Rules

It is clear to the Board that both images violated Meta's prohibition on "derogatory sexualized photoshop" under the Bullying and Harassment policy. Both have been edited to show the head or face of a real person with a real or fictional body that is nude or nearly nude. Both cases also contain contextual clues that the content has been AI-generated. The post in the first case (Indian public figure) includes a list of hashtags indicating that it is AI-generated and was posted by an account dedicated to posting these images. The post in the second case (American public figure) was posted on a Facebook group for AI imagery.

The Board agrees with Meta that only the post in the second case, however, violated the Adult Nudity and Sexual Activity policy, as it depicts the woman having her breast squeezed. The other image does not violate this policy in its current form because it is not a close-up shot of nude buttocks as defined by the company. The Board notes that this means under the current policy, similar images that do not contain obvious contextual clues indicating that they are AI-generated would not be removed. The impact of this on victims is discussed in section 5.2 below.

5.2 Compliance With Meta's Human Rights Responsibilities

In both the first case (Indian public figure) and the second case (American public figure), the Board finds that removal of the content from Meta's platforms is consistent with Meta's human rights responsibilities.

Freedom of Expression (Article 19 ICCPR)

Article 19 of the ICCPR provides for broad protection of expression, including expression that may be considered "deeply offensive" (General Comment 34, para. 11, see also para. 17 of the 2019 report of the UN Special Rapporteur on freedom of expression, A/74/486). When restrictions on expression are imposed by a state, they must meet the requirements of legality, legitimate aim, and necessity and proportionality (Article 19, para. 3, ICCPR, General Comment 34, para. 34). These requirements are often referred to as the "three-part test." The Board uses this framework to interpret Meta's human rights responsibilities in line with the UN Guiding Principles on Business and Human Rights, which Meta itself has committed to in its Corporate Human Rights Policy. The Board does this both in relation to the individual content decision under review and what this says about Meta's broader approach to content governance. As the UN Special Rapporteur on freedom of expression has stated, although "companies do not have the obligations of Governments, their impact is of a sort that requires them to assess the same kind of questions about protecting their users' right to freedom of expression," (A/74/486, para. 41).

I. Legality (Clarity and Accessibility of the Rules)

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

The Board finds that although in this context the term "derogatory sexualized photoshop" should have been clear to the users posting these pieces of content, it is not generally sufficiently clear to users. In response to the Board's question on the meaning of this term, Meta stated that: "'Derogatory sexualized photoshop or drawings' refers to manipulated images that are sexualized in ways that are likely to be unwanted by the target and thus perceived as derogatory (for example, combining a real person's head with a nude or nearly nude body)." The Board notes that a term such as "non-consensual" would be a clearer descriptor than "derogatory" to convey the idea of unwanted sexualized manipulations to images.

Moreover, the Board finds that term "photoshop" in the prohibition on "derogatory sexualized photoshop" is dated and too narrow to cover the array of media manipulation techniques available to users, particularly those powered by generative AI. While the term "photoshop" no longer necessarily implies the use of a particular editing software, it still commonly refers to the manual editing of images using digital tools. By contrast, much of the non-consensual sexualized imagery spread online today is created using generative AI models that either automatically edit existing images or create entirely new ones. Meta should ensure that its prohibition on derogatory sexualized content covers this broader array of editing techniques, in a way that is clear to both users and the company's moderators.

The Board also finds that the policy lines prohibiting these images would be more appropriately located in the Adult Sexual Exploitation Community Standard rather than the Bullying and Harassment Community Standard. The rules need to be intuitive to make it easy for users to understand what is prohibited and why. This is particularly important in cases where the content would be compliant with the rules if the image had been consensually made and shared, such as in the first case (Indian public figure). If a user looks at the images in both these cases, they are unlikely to see them as an issue of Bullying and Harassment.

A public comment from the RATI Foundation for Social Change (an Indian NGO that assists victims of online and offline sexual violence), stated that although one of the ways it assists victims is by helping to get AI-generated sexual images removed from Meta's platforms, it had never heard of the prohibition on "derogatory sexualized photoshop" and had never reported an AI explicit image under Bullying and Harassment. Instead, it reported such images under other policies such as Adult Nudity and Sexual Activity, Child Exploitation and Adult Sexual Exploitation (see PC-27032).

Including this prohibition in the Bullying and Harassment Community Standard presumes that users are posting these images to harass people. However, this may not accurately reflect why a given user has posted an AI-generated explicit image. This is confusing for all users, from the people posting this content to the people reporting it. External research commissioned by the Board shows that users post deepfake intimate imagery for a number of reasons that may not involve an express intent to bully or harass. While harassment and trolling are two of them, users are often also motivated by a desire to build an audience on the platform, monetize their page or direct users to off-platform sites, such as pornography sites and services, or clickbait websites. A study by Powell et. al from 2020 also found that perpetrators of image-based sexual abuse often report motivations of it being "fun" or to "flirt" as well as to "trade the images." The policy lines prohibiting these images would be clearer if the focus was on the lack of consent and the harms of proliferation of such content, rather than the impact of direct attacks implied by a Bullying and Harassment designation.

The Adult Sexual Exploitation policy would therefore be a clearer and more logical place to include these prohibitions. This policy focuses on images shared with a lack of consent and contains the prohibition on non-consensual intimate image sharing (NCII), which is clearly a very similar issue. Meta should also consider renaming this policy to something more detailed and clearer to users, such as "Non-Consensual Sexual Content."

II. Legitimate Aim

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

The Human Rights Committee has interpreted the term "rights" to include human rights as recognized in the ICCPR and more generally in international human rights law (General Comment 34, at para. 28).

Meta's decision to prohibit deepfake intimate imagery on the platform seeks to protect the rights to physical and mental health, as this content is extremely harmful to victims (Article 12 ICESCR); freedom from discrimination, as there is overwhelming evidence showing that this content disproportionately affects women and girls (Article 2 ICCPR and ICESCR); and the right to privacy, as it affects the ability of people to maintain a private life and authorize how images of themselves are created and released (Article 17 ICPPR).

The Board concludes that platform restrictions on deepfake intimate imagery are designed to protect individuals from the creation and dissemination of sexual images made without their consent – and the resulting harms of such images to victims and their rights. This represents a legitimate aim for restricting this content.

III. Necessity and Proportionality

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

The severity of harms in this content:

The Board finds that prohibition and removal are necessary and proportionate measures to protect the rights of people impacted by this content. The harms caused by this content are severe for the people depicted in it. Their rights to privacy and the protection from mental and physical harm are undermined by the non-consensual use of their image to create other sexualized images.

Given the severity of these harms, removal of the content is the only effective means available to protect victims – there are no less intrusive measures that would be sufficient. In the Altered Video of President Biden decision, the Board recommended labeling of manipulated content as a means to prevent users being misled about the authenticity of content. Labeling would not, however, be sufficient to address the harm here, as it stems from the sharing and viewing of the image itself, not solely from people being misled as to its authenticity.

As the vast majority of the people depicted in these images are women or girls, this type of content also has discriminatory impacts and is a highly gendered harm (see PC-27045).

The application of MMS banks to this content:

The Board also considered Meta's use of MMS banks. The image in the second case (American public figure) had already been added to the MMS bank, but the image in the first (Indian public figure) was not added until after the Board asked Meta why this was the case. Meta stated in its response that it relied on media reports to indicate that the second case's image was circulating on social media and that "banking was necessary in order to address the broader issue of the proliferation of this content." It went on to say that there were no media signals in the first case.

The Board highlights that there may be many victims of non-consensual deepfake intimate imagery whose images are shared numerous times on platforms. However, they do not have a public profile and are forced to either accept the proliferation of their non-consensual depictions or search and report every instance, which would be very resource-intensive and traumatizing. A public comment from Witness urges Meta to "avoid placing the burden of reporting on victims, including repeated reporting of the same content," (see PC-27095). The Board reiterates this concern, especially given the impacts on victims in regions or communities with limited media literacy. Meta has stated that one of its signals of lack of consent under the Adult Sexual Exploitation policy is media reports of leaks of NCII images. While this can be useful information for content concerning public figures, it is important that Meta not be over-reliant on this signal as it is not a helpful signal for content concerning private individuals, who will likely not be subject to media reporting. Meta also needs to rely on signals that help identify non-consensual depictions of private individuals.

In determining the proportionality of measures applied by Meta, the Board also discussed the act of sanctioning users – in particular, whether everyone (not just the first user) who shared these images should be given a strike. A public comment from RATI Foundation for Social Change stated that, in their experience of assisting victims of deepfake intimate imagery, “many of these videos are posted in collaboration with another account. However, when the post is actioned only one of the accounts is penalized. The other account which is an alt account of the offender survives and it resumes posting.” It also stated that it saw many copies of the same video, which seems to indicate that MMS banks would be useful in addressing this content (see PC-27032). Of course, as the Digital Rights Foundation notes, MMS banks are “restricted by the database of known images” and will always have a more limited utility against AI-generated images as new ones can be so easily created (see PC-27072). MMS banks, therefore, can only be one tool in Meta’s arsenal to combat deepfake intimate imagery. While applying strikes in every instance could make enforcement more effective, it could also lead to users being penalized in situations where this is not justified, such as sharing images they do not know are AI-generated or otherwise non-consensual. The Board acknowledges this tension. Meta shared with the Board that the MMS bank in this case was not configured to apply strikes due to the risk of over-enforcement. However, in some circumstances, this has changed and users are now able to appeal these decisions. Given this change in circumstances, the Board prompts Meta to reconsider whether applying strikes may be justified.

The artificial distinction between non-consensual intimate image sharing and deepfake intimate imagery:

Finally, the Board considered whether non-consensual intimate image sharing (NCII) and deepfake intimate imagery should be treated separately within Meta's policies. When asked about the possibility of moving the prohibition on derogatory sexualized photoshopping (which, as discussed in the Legality section above, would be better described with a more accurate term) to the Adult Sexual Exploitation Policy, Meta told the Board that the two content categories are very different because the rules on NCII enforcement require a signal of a lack of consent (such as a vengeful statement or media reports of a leak), whereas the rules on derogatory sexualized photoshopping do not. However, this is an enforcement choice that could theoretically be remedied by considering context indicating that the nude or sexualized aspects of the content are AI-generated, photoshopped or otherwise manipulated to be a signal of non-consent, and specifying that such content need not be "non-commercial or produced in a private setting" to violate the policy.

There is already a significant overlap between the policies that may not be clear to users. Meta stated in its response to the Board's questions that, at the time of enforcing the content in these cases, its definition of intimate imagery for the Adult Sexual Exploitation policy was internally defined as (i) screenshots of private sexual conversations and (ii) imagery of one or more people in a private setting, including manipulated imagery that contain nudity, near nudity, or people engaged in sexual activity.

Creating a presumption that AI-generated sexual images are non-consensual may occasionally lead to an image being removed that was consensually made. The Board is deeply concerned about the over-enforcement of allowable nudity and near-nudity, as demonstrated by the Breast Cancer Symptoms and Nudity case, Gender Identity and Nudity cases, and Breast Self-Exam and Testicular Cancer Self-Check Infographics summary decisions. However, in the case of sexualized deepfakes, this presumption has already been underlying Meta's enforcement of derogatory sexualized photoshopping, as the company presumes that all sexualization covered by this policy and created through AI or photoshopping is unwanted. It is inevitable that not all AI-generated content will be caught by this new policy line (just as it is not caught now), but by combining the two categories of non-consensual content, Meta can leverage its successes at combatting NCII and use aspects of its approach to assessing consent to reduce deepfake intimate imagery on its platforms.

The Board also explored whether, in order to provide better protection for those whose rights are impacted by this type of content, Meta should alter its approach to the prohibitions on NCII and derogatory sexualized photoshop to start with a presumption that such imagery is non-consensual, instead of the current approach of presuming imagery is consensual and requiring signals of non-consent to remove them. After assessing the feasibility and impact of this proposed approach, however, the Board concluded that such an approach risked significantly over-enforcing against non-violating content and would not currently be operationally feasible in the context of automated tools that Meta relies on for enforcement.

Access to Remedy

The Board is concerned by the appeals that were auto-closed in the first case. Both the original report and the appeal against Meta's decision to keep the content on the platform were auto-closed. Meta informed the Board that "content reported for any violation type (with the exception of Child Sexual Abuse Material) is eligible for auto-close automation if our technology does not detect a high likelihood of a violation and it is not reviewed within 48 hours."

Users may be unaware of the auto-closing process and the fact that when they submit content for appeal, it may never actually be reviewed. Meanwhile, as in the first case, victims and others seeking to remove deepfake intimate imagery may report the content but are denied any actual review. When they then appeal that decision, they can find themselves in the same position, with the same auto-closing process happening again.

Many of the public comments received in this case criticized the use of the auto-closing of appeals for image-based sexual abuse. The damage caused by these images is so severe that even waiting 48 hours for a review can be harmful. The American Sunlight Project, which gave the example of deepfake intimate imagery targeting female politicians during elections, states, “such content could receive hundreds of thousands of views, be reported on in national press, and sink the public perception of a political candidate, putting her on uneven footing when compared with her opponents. In some countries, including India, where this case took place, it could even endanger her life,” (see PC-27058). A related point was made by the Centre for Protecting Women Online, which cautioned that the harm of these images in an election will be particularly severe “in contexts where digital literacy in the general population is low and where the influence of messages and images posted on social media is highly likely to influence voters as the almost exclusive source of news and information,” (PC-27088). Of course, regardless of the public or private status of victims, a delay in removing these images severely undermines their privacy and can be catastrophic. The Board considered whether content that causes such severe harms to its victims (through both deepfake and NCII) should be exempt from the auto-closing process. The Board acknowledges the challenges of at-scale content moderation, and the need to rely on automated processes to manage content flagged for review, but it is concerned about the severity of harm that may result from its use in policy areas such as this one. The Board does not have sufficient information on the use of auto-closing across all Meta’s policies to make a recommendation on the use of auto-closing within Meta’s broader enforcement systems, but considers it an issue that may have significant human rights impacts that require careful risk assessment and mitigation.

6. The Oversight Board’s Decision

The Oversight Board overturns Meta’s original decision to leave up the content in the first case (Indian public figure), requiring the post to be removed, and upholds Meta's decision to take down the content in the second case (American public figure).

7. Recommendations

A. Content Policy

1. To increase certainty for users and combine its policies on non-consensual content, Meta should move the prohibition on "derogatory sexualized photoshop" into the Adult Sexual Exploitation Community Standard.

   The Board will consider this recommendation implemented when this section is removed from the Bullying and Harassment policy and is included in the publicly available Adult Sexual Exploitation policy.

2. To increase certainty for users, Meta should change the word "derogatory" in the prohibition on "derogatory sexualized photoshop" to "non-consensual."

   The Board will consider this recommendation implemented when the word "non-consensual" replaces the word "derogatory" in the prohibition on derogatory sexualized content in the publicly available Community Standards.

3. To increase certainty for users and ensure that its policies address a wide range of media editing and generation techniques, Meta should replace the word "photoshop" in the prohibition on "derogatory sexualized photoshop" to a more generalized term for manipulated media.

   The Board will consider this recommendation implemented when the word "photoshop" is removed from the prohibition on "derogatory sexualized" content and replaced with a more generalized term, such as "manipulated media."

4. To harmonize its policies on non-consensual content and help ensure violating content is removed, Meta should add a new signal for lack of consent in the Adult Sexual Exploitation Policy: context that content is AI-generated or manipulated. For content with this specific context, the policy should also specify that it need not be "non-commercial or produced in a private setting" to be violating.

   The Board will consider this recommendation implemented when both the public-facing and private internal guidelines are updated to reflect this change.

Procedural Note:

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

- Under its Charter, the Oversight Board may review appeals from users whose content Meta removed, appeals from users who reported content that Meta left up, and decisions that Meta refers to it (Charter Article 2, Section 1). The Board has binding authority to uphold or overturn Meta's content decisions (Charter Article 3, Section 5; Charter Article 4). The Board may issue non-binding recommendations that Meta is required to respond to (Charter Article 3, Section 4; Article 4). Where Meta commits to act on recommendations, the Board monitors their implementation.

- For this case decision, independent research was commissioned on behalf of the Board. The Board was assisted by Duco Advisors, an advisory firm focusing on the intersection of geopolitics, trust and safety, and technology. Memetica, a digital investigations group providing risk advisory and threat intelligence services to mitigate online harms, also provided research.