Reported AI-Generated Sexualized Video

2026-021–IG-RA

Summary

The Oversight Board has overturned Meta's decision to leave up on Instagram a reportedly AI-generated video impersonating a woman and calls on Meta to strengthen protections for non-public figures who are the targets of sexualized deepfakes.

Why This Matters

The Board's decision and recommendations come at a time when governments, such as in India, the UK and Spain, are setting up new rules for platforms on AI-generated content. The European Union (EU) has reached an agreement to prohibit AI systems that generate non-consensual sexually explicit and intimate content or child sexual abuse material, such as AI 'nudification' apps. Other platforms, such as X, are facing scrutiny over sexually explicit images made by AI chatbots.

It is clear that the scale, speed and sophistication of AI tools have resulted in a proliferation of AI-generated sexualized non-consensual content globally. The spread of sexualized deepfake videos leads to reputational and psychological harm, which disproportionately impacts women and girls, and has a chilling effect on participation in social and political life.

About the Case

In September 2025, an Instagram user posted a reportedly AI-generated eight-second video of a woman adjusting her form-fitting dress and moving her body, with her underwear visible in a few frames.

The next day, Meta's automated system that detects content that may pose harm to individuals and has a high likelihood of virality identified the post, but it was not prioritized for human review. A few days later, two users reported the content, but it remained on the platform. One of the users appealed to Meta, but the post was not reviewed, and the user then appealed to the Board. The user who appealed told the Board that the video was AI-generated and that it impersonated a friend (who had already closed her account) of theirs without the friend's consent.

When the Board brought the case to Meta's attention, the company's subject matter experts reviewed the post and concluded that it did not merit removal under the company's Community Standards but made the post visible only to adults.

Key Findings

The Board finds that the post violates the prohibition against sharing non-consensual intimate imagery (NCII) under Meta's Adult Sexual Exploitation Policy. This is because the post fulfills all three of the policy's criteria for non-consensual intimate imagery – it appears to be non-commercial content in an apparently private setting; the woman depicted is "near nude"; and there is a lack of consent.

Meta said that when the content was flagged, the company had no indication that the individual depicted in the video was "a real person" because they did not report the content. Among the signals Meta uses for determining lack of consent is a report from the person depicted, in addition to captions, comments or titles that suggest a vengeful context; or reports from independent sources such as law enforcement, media or representatives of survivors of non-consensual intimate imagery.

The Board finds that AI-generated impersonation is non-consensual by default and should be added to the set of signals the company uses to establish lack of consent. Broadening the signals of lack of consent in this way would especially benefit non-public figures who are the targets of non-consensual intimate imagery because it would reduce the burden on victims to report the abuse themselves. Experts the Board consulted during the case said victims benefit most from rapid removal and better ways to flag content that do not rely on self-reporting.

To further reduce this burden, the Board finds that Meta should allow verified accounts from friends and families of those targeted to report violating content on their behalf.

Meta should also improve the process for reporting and appealing non-consensual intimate imagery globally on its platforms. Meta and other large language model developers should embed safeguards in system design and implement content credentials at scale.

The Oversight Board's Decision

The Board overturns Meta's decision to leave up the content and requires that the post be removed.

The Board recommends that Meta:

- Should add a new signal for lack of consent in the Adult Sexual Exploitation policy: context that content is AI-generated sexualized impersonation of real people.
- Allow users to designate "connected accounts," such as trusted friends, family members or associates, that can report on their behalf potential Community

Standards violations involving non-consensual intimate imagery, including impersonations.

- Include AI-generated sexualized impersonation as a separate category in standard content reporting and appeal forms, distinct from "harassment" or "nudity." The reporting forms should be made available globally.

The Board also reiterates the following recommendations from earlier decisions:

- Change the word "derogatory" in the prohibition on "derogatory sexualized photoshop" to "non-consensual" and replace the word "photoshop" with a more generalized term for manipulated media.
- Implement content credentials (as laid out by the Coalition for Content Provenance and Authenticity (C2PA)) at scale and ensure that they are clearly and consistently visible and accessible to users whenever the provenance details are available.

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

Full Case Decision

1. Case Description and Background

The scale, speed and sophistication of AI tools have resulted in the proliferation of AI-generated sexualized non-consensual content. In 2023, a report from the online security platform Security Hero found that 98% of deepfake videos online were "deepfake pornography" and 99% targeted women. A 2024 study by the non-governmental organization (NGO) Internet Matters found that 99% of nude deepfakes featured women and girls.

In a February 2026 joint statement, 61 data protection and privacy authorities from around the world warned that "recent developments, particularly AI image and video generation integrated into widely accessible social media platforms, have enabled the creation of non-consensual intimate imagery, defamatory depictions and other harmful content featuring real individuals." According to the UK news group the Guardian, as AI models create images by learning from and replicating elements of existing images, "nudification apps are thought to be trained on vast datasets of mostly female images because they tend to work most effectively on women's bodies."

A 2025 survey with over 16,000 adults in 10 countries revealed that women reported greater harms and negative impacts from image-based sexual abuse than men. Research shows that image-based non-consensual sexual abuse leads to reputational and psychological harm – such as distress, anxiety and trauma – harassment and humiliation, impacts on identity, autonomy and integrity, and a chilling effect on participation in social and political life.

Public comments submitted to the Board underline the importance of considering cultural and local contextual differences in defining and enforcing against AI-generated non-consensual sexualized content, including impersonation (see PC-32490 and PC-32481). They provide examples of AI-generated non-consensual intimate imagery involving women who are both public figures and non-public figures. The NGO Digital Rights Foundation Pakistan noted that its Digital Security Helpline has received multiple cases of AI-generated non-consensual image-based abuse, including cases where Pakistani women living in Italy and the UK have been targeted by abuse and later were victims of honor-based killing "either there or after being forced to come to Pakistan" (PC-32481).

A significant number of jurisdictions have moved to regulate and criminalize non-consensual deepfakes or AI-generated sexualized imagery. Spain has proposed legislation aimed at curbing AI deepfakes and reinforcing consent rules for using

images, voices and likenesses. Mexico passed the “Olimpia Act” focused on legislative amendments addressing digital violence, which includes non-consensual intimate imagery. Peru and Colombia have passed laws that make the use of deepfakes an aggravating factor to a crime. In Canada, British Columbia’s Intimate Images Protection Act focuses on non-consensual distribution of intimate imagery.

In the U.S., the Take it Down Act, signed into law in 2025, makes it a crime to publish or threaten to publish non-consensual intimate imagery, including AI-generated deepfakes, and requires platforms to remove such content within 48 hours of notification. The Indian government requires platforms to remove non-consensual sexual content within 24 hours after a complaint is made. The UK has implemented a 48-hour removal window for non-consensual intimate imagery.

The European Union (EU) has reached an agreement to prohibit AI systems that generate non-consensual sexually explicit and intimate content or child sexual abuse material, such as AI ‘nudification' apps.

This case raises emblematic policy and enforcement considerations in the context of AI-generated sexualized imagery, including sexualized impersonation. In 2024, the Board discussed two cases of explicit AI images that resembled female public figures from India and the U.S. (Explicit AI Images of Female Public Figures). Building on that decision, the present case considers issues concerning AI-generated sexualized impersonation, particularly involving non-public figures.

In September 2025, an Instagram user posted a reportedly AI-generated eight-second video of a woman in a form-fitting dress. In the video, the woman is adjusting her dress and moving her body, with her underwear visible in a few frames.

The next day, Meta's automated system that detects and prioritizes content that may pose harm to individuals, especially content with a high likelihood of virality, identified the post. However, the report was not prioritized for human review.

A few days later, two users reported the content for pornography, but their reports were not prioritized for human review either, and the content remained on Instagram. One of the reporting users appealed to Meta to take the content down, but again, the post was not prioritized for human review. The user then appealed to the Board.

When the Board brought the case to Meta's attention, Meta's subject matter experts reviewed the post and concluded that it did not violate the company's Community Standards. Therefore, Meta did not remove the content, but it did determine that it should only be visible to adults under the Adult Nudity and Sexual Activity policy.

2. User Submissions

In appealing to the Board, the reporting user stated that the video was AI-generated and impersonated a woman, one of the reporting user's friends, without her consent, therefore "damaging her reputation." The reporting user added that their friend had already closed her account, but the posting user continued to make sexualizing content using their friend's image.

3. Meta's Content Policies and Submissions

Content Assessment

Meta found that the post does not violate any Community Standard. Meta did not consider the post to violate the Adult Nudity and Sexual Activity policy, which prohibits photorealistic or digital imagery of adult nudity, or photorealistic or digital videos that focus on crotch, female breasts or buttocks recorded without the awareness of the person(s) depicted in them. However, under this policy, the company restricts visibility of "photorealistic or digital imagery of persons where crotch, buttock or female breast(s) are the focus of the image" to users over 18. Because the depicted woman draws attention to her crotch area and underwear by moving her legs and adjusting the clothing, Meta determined the content falls under this policy line and should only be visible to adults.

Meta also did not consider that the post violated the Adult Sexual Exploitation Community Standard, which prohibits "sharing, threatening, stating an intent to share, offering or asking for non-consensual intimate imagery (NCII) that fulfils all of the three following conditions:

a. Imagery is non-commercial and produced in a private setting.
b. Person in the imagery is (near) nude, engaged in sexual activity or in a sexually suggestive pose (this includes digitally created or AI-generated imagery).
c. Lack of consent to share the imagery is indicated by meeting any of the signals:
    i. Vengeful context (such as caption, comments or Page title).
    ii. Independent sources such as law enforcement records, media reports (such as leaks of images confirmed by media) or representatives of a survivor of NCII.

iii. Report from a person depicted in the image or who shares the same name as the person depicted in the image."

According to Meta, at the time of escalated review, the company had no indications, such as user reports, suggesting that the depicted individual was "a real person." Meta noted that had the depicted individual reported the content, it would have violated the Adult Sexual Exploitation policy. This is so because the depicted individual wears lingerie or sleepwear, which constitutes near nudity in violation of the sharing non-consensual intimate imagery policy line. To Meta, the report would have served as a clear signal of non-consent.

Meta also did not consider that the post violates the Bullying and Harassment policy that, under its Tier 1 [universal protections for everyone], prohibits "derogatory sexualized photoshop or drawings." This includes content that has been altered to combine a real person's head or face with a body, real or fictional, that is either sexually explicit (nude, near nude or engaged in sexual activity) or posed in a sexually suggestive manner. Meta stated that the individual in the video is not depicted in a sexually explicit manner, where they are nude or nearly nude. To Meta, the individual's poses similarly do not constitute sexually suggestive poses as defined under the Adult Nudity and Sexual Activity policy, which include spreading legs to expose the genitalia area or bending over while wearing underwear or see-through clothing.

Because the three policies referred to above use the term "near nudity," the Board asked Meta why the post only constituted near nudity in violation of the sharing non-consensual intimate imagery policy line. In response, Meta explained that the key distinction between near nudity and nudity is how the private parts are covered: in the case of near nudity, they are covered only by an opaque object (e.g., unworn clothing), a digital overlay or a human body, making them partially visible but not fully exposed. Conversely, full nudity does not involve such a covering. According to Meta, the three policies share the same baseline definition of near nudity, namely, a "human being that

is not fully clothed in specific ways" to partially reveal pubic hair, genitalia, female nipples and/or buttocks, when covered by objects, digital overlays or see-through clothing. However, according to Meta, only the Adult Sexual Exploitation policy builds on that shared definition by including additional indicators of near nudity for the purposes of identifying potential non-consensual intimate imagery. For Meta, this broader approach reflects how non-consensual intimate imagery abuse may appear on its platforms, i.e., manifesting as intimate imagery, featuring individuals in private settings wearing lingerie or other intimate apparel, which goes beyond the types of nudity and near nudity defined under the other policies. Meta further explained that while the baseline definition is maintained for the enforcement of the Adult Nudity and Sexual Activity policy to avoid removing benign content, without the expanded definition, non-consensual intimate imagery could potentially evade enforcement.

With regard to the Bullying and Harassment policy, the company explained that derogatory sexualized photoshop or drawings are considered inherently derogatory, while non-consensual intimate imagery typically involves real people captured in private or intimate settings and is shared without their consent. According to Meta, given that these represent two distinct abuses, the broadened definition of near nudity under the Adult Sexual Exploitation policy is not applicable to the derogatory sexualized photoshop or drawings policy line under the Bullying and Harassment policy.

Meta further explained that, unlike the "near nude" standard, the definition of "sexually suggestive poses," referred to above, remains the same across all policies.

In response to the Board's questions, Meta stated that since the company did not detect International Press Telecommunications Council (IPTC) or C2PA metadata at scale in videos at the time the content was posted, and the posting user did not self-disclose that the video was created or altered using AI, the content in this case was not labelled as AI-generated or manipulated. Meta noted that the company requires users to

disclose, through the AI-disclosure tool, whenever they post organic content with photorealistic video or realistic-sounding audio that was digitally created or altered, and the users may receive penalties for failure to self-disclose (see also AI-Generated Video in Israel-Iran Conflict decision).

Determining Consent

Meta told the Board that the requirement for determining non-consent varies across its Community Standards. Regarding sexualized content, some policies, such as the Adult Sexual Exploitation, require signals of non-consent for removal. Other policies, such as the Adult Nudity and Sexual Activity or certain policy lines under Bullying and Harassment, will be considered violative regardless of the consent of the depicted or targeted person. However, for other Bullying and Harassment policy lines (which are not applicable to this case), signals that the content is unwanted are required content to be removed, and these are evaluated as a proxy for non-consent.

Specifically, under the Adult Sexual Exploitation policy, for non-commercial non-consensual intimate imagery content, including when it's AI-generated, to be removed, signals of non-consent are required. Meta explained that the company determines if sexually explicit or suggestive imagery has been shared non-consensually based on the context of the violation.

The company added that it does not require a direct report from the depicted individual to remove non-consensual intimate imagery if there are other credible indicators of non-consent. These include third-party reports from law enforcement, media or trusted partners, as well as contextual cues within the content itself (e.g., in the caption, comments or page title) suggesting that the imagery was shared in a "vengeful or sensationalist manner."

Conversely, violations of the Adult Nudity and Sexual Activity policy are removed regardless of the consent of the depicted individual. Similarly, violations involving derogatory sexualized photoshop or drawings under the Bullying and Harassment policy are considered inherently unwanted, and therefore, no self-reporting is required.

Finally, for other Bullying and Harassment policy lines, the company evaluates whether content is unwanted through self-reporting or contextual signals. For example, self-reporting applies to the "Tier 3" unwanted manipulated imagery rule (which does not need to be sexual or intimate in nature), as it is not evidently derogatory or offensive. Whereas for the "Tier 1" rule regarding unwanted contact that is sexually harassing, no self-reporting is required, but Meta considers several signals to determine that the content depicts behavior that is unwanted, including whether the recipient self-reported, expressed discomfort, asked for the contact to stop or took on-platform actions such as reporting or blocking.

Third-Party Reporting

Meta informed the Board that signals of non-consent apply equally to non-consensual intimate imagery involving public figures and non-public figures. The company explained that at scale, it does not rely on reports from other third parties who claim to be acting on behalf of the depicted individual, as the company cannot reliably verify whether those reports are authorized or otherwise credible. To Meta, third-party reporting, other than from trusted flaggers or partners, carries the risk of false non-consensual intimate imagery claims on content that may have been shared by consenting adults, potentially leading to overenforcement.

Meta added that the company allows both depicted persons and third parties to report impersonation on Facebook and Instagram. Meta's Authentic Identity Representation policy prohibits accounts that impersonate another person or entity by using their image, name or likeness. Meta noted that it prohibits accounts that deliberately mislead

others about their identity through, for example, repeated or significant changes to identity details, or misleading profile information. The company explained that this policy focuses on account-level rather than content-level impersonation, as users may share others' content without explicit intent to deceive (e.g., posting photos of friends). Meta considers that enforcing at the content level in these contexts could lead to overenforcement. The company added that impersonation at the content level is separately enforced on escalation under the Fraud, Scams, and Deceptive Practices policy, e.g., for content that makes unauthorized use of an image of a famous person in an attempt to scam or defraud.

The Board asked questions on policy and enforcement considerations for determining lack of consent at scale, particularly for content involving non-public figures, avenues for third-party reports on behalf of depicted individuals, beyond law enforcement, media and trusted partner reports, penalties for failure to self-disclose AI use in video or audio generation, and the interplay between several relevant Community Standards. Meta responded to all the questions.

4. Public Comments

The Board received 13 public comments that met the terms for submission. Three of the comments were submitted from Europe, three from United States and Canada, two from Asia Pacific and Oceania, two from Central and South Asia, two from Sub-Saharan Africa, and one from Latin America and the Caribbean. To read public comments submitted with consent to publish, click here.

The submissions covered the following themes: implications of a standardized definition of sexualized content, given cultural and local differences; need to streamline policy approaches to non-consensual intimate imagery; age-gating as an inappropriate measure to address non-consensual intimate imagery; importance and benefits of allowing third parties to report on behalf of survivors of non-consensual intimate imagery; recommendations on determining lack of consent at scale, improving enforcement of non-consensual intimate imagery and implementing product-level safeguards.

In March 2026, as part of ongoing stakeholder engagement, the Board consulted with representatives of advocacy organizations, academics, inter-governmental organizations and other experts on the issue of moderating AI-generated non-consensual sexualized content. The participants provided insights on the use and proliferation of AI-generated sexualized content, including through impersonations. They highlighted the importance of treating non-consensual AI sexualized impersonation as non-consensual sexual content, not merely as a question of adult nudity. They suggested that social media companies should adopt a presumption of non-consent, given the risk of serious harm from non-consensual intimate imagery, including harassment, exploitation and reputational harm. Some participants noted that non-public figures should be afforded a higher level of protection, as they often have less access to media and other resources to report the abuse. Others underlined

the importance of directly involving survivors of non-consensual intimate imagery abuse in policy and enforcement design processes.

5. Oversight Board Analysis

The Board selected this case to assess Meta's approach to AI-generated sexualized content, particularly in the context where a person's image or likeness is used in a sexually explicit or suggestive manner without their consent. The case is relevant to two of the Board's seven strategic priorities, Automation and AI, and Gender.

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

The Board finds that the post violates Meta's prohibition against sharing non-consensual intimate imagery under its Adult Sexual Exploitation policy. This post represents an AI-generated sexualized impersonation as a type of image-based sexual abuse where a person's image or likeness is used in a sexually suggestive or explicit manner without the person's consent.

To violate Meta's rule on non-consensual intimate imagery, the content should meet the conditions set out in the Adult Sexual Exploitation policy, and the content in this case fulfills all three. Firstly, the video appears to be of a non-commercial nature and seems to be produced in a private setting. Secondly, the depicted woman is wearing lingerie, which satisfies the definition of near nudity under the policy.

With respect to the third condition, the Board notes that AI-generated impersonation is non-consensual by default. Although Meta argues that the depicted woman is not "a real person," in their appeal to the Board, the user stated that the video was AI-generated and impersonated one of their friends, who had closed her account, presumably due to the posting user's continued use of her image to make sexualized content. The Board is concerned that the company concludes that the depicted person is not "a real person" based solely on the lack of a user report. This illustrates a practical gap in determining consent to enforce Meta's rules on non-consensual intimate imagery and AI-generated sexualized impersonation content, especially involving non-public figures.

Although Meta's October 2025 update to the Adult Sexual Exploitation policy to include AI-generated imagery is a necessary step forward, given the proliferation of AI-generated sexualized impersonation content worldwide, additional steps are required. Given the severe harms of AI-generated sexualized impersonations, the company should default to prioritizing safety.

To this end, and in line with previous recommendations made by the Board (see e.g., Explicit AI Images of Female Public Figures, recommendation no. 4), Meta should broaden the scope of signals of lack of consent, especially for non-public figures. While the Board believes that, in line with expert and stakeholder feedback received, the broader approach previously recommended (considering AI-generation as a contextual signal of lack of consent for non-consensual intimate imagery under the Adult Sexual Exploitation policy) better addresses the harms caused by such content, Meta did not meaningfully implement the recommendation.

Therefore, on the policy level, at the very least, the company should consider explicitly listing AI-generated sexualized impersonation of real people as a contextual signal of non-consensual intimate imagery to avoid potential policy and enforcement gaps. This

would require a contextual assessment of signals in the content (e.g., post caption, page title or comments), similar to, for example, assessment of vengeful context as a signal of lack of consent under the Adult Sexual Exploitation policy. Improved reporting flows and reports from designated connecting accounts, as discussed below, would also provide relevant signals for these assessments. At the same time, the absence of these contextual signals, coupled with user self-disclosure of utilization of AI tools to create sexualized imagery, may support the conclusion that the content does not depict a real person, and therefore, is not an impersonation. This approach is sufficiently narrowly drawn to prevent overenforcement but captures violative AI-generated sexualized impersonation imagery.

The Board is also concerned that Meta's current approach primarily focuses on accounts that impersonate another person or entity under the Authentic Identity Representation policy, and impersonation of public figures at the content level under the Fraud, Scams and Deceptive Practices policy. This approach does not account for the well-documented, acute harms of AI-generated sexualized impersonation imagery on non-public figures.

II. Enforcement Action

This case raises two distinct and interrelated concerns about non-consensual intimate imagery enforcement. The first regards the limitations on the number of entities other than the person depicted who can signal lack of consent under the Adult Sexual Exploitation policy; and the second is about the effectiveness of current reporting and appeal flows. Changes to both mechanisms would lead to improved enforcement against non-consensual intimate imagery, including AI-generated sexualized impersonation. This case also illustrates the need to embed safeguards in system design and implement content credentials at scale.

Designating Connected Accounts

Meta's responses to the Board's questions suggest that, currently, the only effective pathway for non-public figures to establish lack of consent is to self-report.

Non-public figures would hardly directly benefit from law enforcement and media reports submitted to Meta, as, e.g., they would have to file a separate complaint to law enforcement and depend on the respective public institution's decision to act on it. Additionally, given the proliferation of AI-generated sexualized content, including impersonation, reports from trusted partners and representatives of survivors of non-consensual intimate imagery abuse would likely offer a limited opportunity to ascertain lack of consent. In this regard, while it is a positive step to include reports from representatives of survivors of non-consensual intimate imagery abuse as an additional source to establish non-consent, these are closer to trusted partner reports and are likely to be insufficient for first-time victims.

Public comments submitted to the Board highlight the impact of Meta's current approach to determine lack of consent at scale for victims of image-based abuse, including impersonation (see PC–32480, PC-32489, and PC-32484). The submissions argue that this approach puts the burden on victims and includes the risk of revictimization and retraumatization. Additionally, self-reporting may also be challenging when impacted users do not have active accounts on Meta's platforms to monitor and report violative content, as in this case. Experts consulted by the Board note that the most effective systems to report on the use of image and likeness in AI-generated content are those that reduce the burden on victims, enable batch or cluster reporting, allowing reporting on multiple posts at the same time, and trigger rapid removal.

To address some of these enforcement gaps, in Explicit AI Images of Female Public Figures, the Board recommended that if the nude or sexualized aspects of content are AI-generated, photoshopped or otherwise manipulated, even if such content is not

"non-commercial or produced in a private setting," this should be viewed as a signal of non-consent. Although in response, Meta updated the definition of non-consensual intimate imagery to include digitally created or AI-generated imagery, the Board does not consider the recommendation to be implemented – rather, reframed – as the company still does not consider AI-generation as a signal of lack of consent.

Therefore, in the face of surging AI-generated sexualized impersonation imagery and to reduce the burden on victims of such abuse, additional mechanisms to ascertain lack of consent are called for. In this regard, Meta should allow users to designate connected accounts to report potential Community Standards violations involving non-consensual intimate imagery, including impersonations, with their agreement and on their behalf. To address similar concerns in a different context, the Board has already made recommendations on designating third parties to submit reports under Bullying and Harassment policy lines that require self-reporting (see Gender Identity Debate Videos, recommendation no. 3). Some Board Members believe that Meta should also consider providing similar reporting opportunities for third parties who have a Meta account and recognize individuals depicted in the content who do not have an account (as in the current case), or who have not been designated as a connected account, to report the content.

Reporting and Appeal Flows

Several public comments point to the shortcomings of the existing reporting and appeal flows, underlining the importance of meaningful and efficient review of reports on AI-generated non-consensual intimate imagery, including impersonation (see PC–32481, PC-32489, and PC-32484).

Meta's Help Center page on reporting non-consensual intimate imagery states that users can report non-consensual intimate imagery through existing user reporting channels or through a specialized reporting form, which is currently only available for

residents of the U.S. states of Texas and Florida (while the company notes that it is "in the process of making the specialized form available to all US residents"). The specialized reporting form enlists two reasons for the report: intimate imagery (defined as "nudes, partial nudes or sexually explicit content shared without consent"), and deepfake intimate imagery (defined as "AI-generated or fake sexual content made to look like someone without their consent").

AI-generated non-consensual intimate abuse, including sexualized impersonations, is a global issue (see e.g., PC-32484, PC-32487, PC-32490, PC-32481 and PC-32480). Therefore, all users on Meta's platforms should have easily accessible pathways to report such content globally and have it routed to dedicated review channels. This should include listing AI-generated sexualized impersonation as a separate category in regular report and appeal forms, distinct from harassment or nudity. Since the specialized report form is only accessible to certain users in the U.S., these changes should be adopted in regular report and appeal forms, where global users report content on Meta's platforms. Experts consulted by the Board note that reliance on generic categories for reporting, such as nudity or harassment, often leads to inconsistent outcomes.

Additional Enforcement Mechanisms

At the same time, the Board reiterates its recommendation in AI-Generated Video in Israel-Iran Conflict, which called on Meta to implement content credentials (as laid out by the C2PA) at scale and ensure that they are clearly and consistently visible and accessible to users whenever the provenance details are available, beyond back-end systems and internal detection. This would ensure that both Meta and users have additional reliable signals that the sexualizing content may be AI-generated or manipulated. It would also serve as a new measure to enforce the signal of being non-consensual or unwanted. This is particularly important for non-public figures who are often subject to harassment, reputational harm and even threats to their physical safety

that can be life-disrupting, stigmatizing and isolating. Although harms to both public and non-public figures are severe, public figures have more meaningful reporting pathways than non-public figures. In response to this recommendation, Meta committed to enhancing how it presents content provenance details to users, and acknowledged that "there is meaningful work ahead in scaling interventions across other content types so users can better distinguish between content that was fully created by AI and content that was only partially edited with AI tools, and how to ensure a consistent experience across Facebook, Instagram and Threads."

Expert reports and submissions to the Board also highlight the importance of embedding safeguards in system design. A 2023 report by social network analysis company Graphika notes that AI-generated non-consensual intimate imagery "has moved from a custom service available on niche internet forums to an automated and scaled online business that leverages a myriad of resources to monetize and market its services." According to a 2025 report by NGO AI Forensics, over 3,000 pornographic ads were approved and circulated through Meta's advertising system in the previous year, which also featured AI-generated pornographic media, such as images, video and audio. Therefore, system-level safeguards should include auditing and enforcing against AI-generated pornographic ads and ensuring safety by design for frontier model training, deployment and governance, taking into consideration impacts on women and girls.

Finally, in assessing this case, the Board also considered scaled solutions to protect users' images and likenesses, such as opt-in features to match and flag a user's facial features to potentially non-consensual uses of their image or likeness (see e.g., PC-32489, and PC-32484). However, such mechanisms were discarded, as their implementation mainly rests on employing facial recognition technologies, which entail significant privacy, surveillance and bias concerns, and a documented disproportionate impact on marginalized communities.

5.2 Compliance With Meta's Human Rights Responsibilities

The Board finds that removal of the content would be consistent with Meta's human rights responsibilities.

Freedom of Expression (Article 19 ICCPR)

Article 19 of the International Covenant on Civil and Political Rights (ICCPR) provides for broad protection of expression, including 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). 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 United Nations (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.

In Explicit AI Images of Female Public Figures, the Board highlighted several concerns around clarity and accessibility of Meta's rules on non-consensual intimate imagery. The Board recommended that Meta:

- House all relevant rules under the Adult Sexual Exploitation policy, moving the "derogatory sexualized photoshop" policy lines from Bullying and Harassment there (recommendation no. 1).
- Change the word "derogatory" in the prohibition on "derogatory sexualized photoshop" to "non-consensual," given that "non-consensual" would be a clearer descriptor than "derogatory" to convey the idea of unwanted sexualized manipulations to images (recommendation no. 2).
- Ensure its policies address a wide range of media editing and generation techniques, by replacing the word "photoshop" in the prohibition on "derogatory sexualized photoshop" to a more generalized term for manipulated media (recommendation no. 3).

Although these recommendations were issued almost two years ago, Meta has not meaningfully implemented them. After updating the Adult Sexual Exploitation policy in 2025, to include digitally created or AI-generated imagery under the non-consensual intimate imagery policy line, Meta considers recommendation no. 1 implemented, while the Board considers it reframed. Regarding recommendation no. 2, Meta has consistently noted that the company continues to consider ways to clarify the Bullying

and Harassment policy, though it does not expect to make changes to it, as specified in the recommendation. Similarly, of recommendation no. 3, Meta noted that the company is still in the process of updating the relevant language in the Bullying and Harassment policy. The Board calls on Meta to fully implement recommendations no. 2 and 3, to ensure clarity to users in the context of its rules on non-consensual intimate imagery, including AI-generated sexualized impersonation.

II. Legitimate Aim

Any restriction on freedom of expression should also pursue one or more of the legitimate aims listed in the ICCPR, which include protecting the rights of others (Article 19, para. 3, ICCPR).

The Board has previously recognized that Meta's restrictions on deepfake intimate imagery are designed to protect individuals from the creation and dissemination of sexual images without their consent, and the resulting harms of such images to victims and their rights (Explicit AI Images of Female Public Figures). Those include 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).

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).

Users' rights to privacy and protection from mental and physical harm are undermined by the non-consensual use of their image to create other sexualized images. AI-generated sexualized impersonation content represents an ongoing trend of loss of control over one's image and identity. Unlike deepfake non-consensual intimate imagery that entails, for example, placing a real person's face on an explicit image or video, AI-generated sexualized impersonation imagery uses a real person's image or likeness without their consent to fabricate a person that resembles them. The Board reiterates that 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 e.g., PC-32480, 32481, 32484, 32487, 32488, 32490; see also Explicit AI Images of Female Public Figures).

Therefore, given the gravity of the harms and that no less intrusive measures would be sufficient, removal of such content, including the post in this case, is a necessary and proportionate measure. As the harm stems from sharing and viewing of such images, not solely from misleading people as to their authenticity, labeling would not be a sufficient measure (see Explicit AI Images of Female Public Figures). Similarly, public submissions to the Board highlight that findings of no violation and age-gating in this case "can become a de facto dismissal of the core allegation around non-consensual sexualized impersonation" (see, e.g., PC-32481). Hence, age-gating AI-generated sexualized impersonation content is not necessary and proportionate, as this measure does not meaningfully stop the further circulation of such content, thereby not preventing reputational harm, harassment and potential physical abuse from occurring. Additionally, since Meta's enforcement mechanisms address harm after detection, rapid removal is crucial to reduce virality and associated harm to impacted people (see also PC-32489).

Several public submissions to the Board further note the importance of considering cultural and local differences in enforcing rules on sexualized content (see, e.g., PC-

32490 and PC-32481). Experts consulted by the Board underline that many platforms adopt Western-centric definitions of sexual harm, focusing mostly on explicit nudity or sexual acts. The experts state that in certain socially conservative communities, "AI-generated videos or images depicting a woman in public with men (or even fully clothed in private), even in non-explicit scenarios, can lead to severe reputational damage, social ostracism, forced marriage, honor-based violence or physical harm" (see also PC-32481). Moreover, Meta recently announced its plan to use more advanced AI for improved content enforcement, noting that these systems can "increase capacity in any language based on need and adapt to understand cultural nuance – including niche subcultures – rapidly changing and regionally specific code words, emoji meanings and slang." The Board reinforces the expert conclusion that effective policy enforcement should evaluate harms in context to account for these local and cultural nuances, including for sexually suggestive content, and looks forward to exploring the opportunities provided by advanced AI systems, through independent and transparent oversight.

In Explicit AI Images of Female Public Figures, the Board acknowledged that creating a presumption that AI-generated sexual images are non-consensual may occasionally lead to an image being removed that was consensually made, raising concerns about overenforcement of allowable nudity and near nudity. However, given the severe harms and the fact that Meta's rules on non-consensual intimate imagery also include AI-generated manipulated intimate imagery, Meta should broaden the scope of signals of lack or consent, especially with regard to non-public figures. The company should, at a minimum, consider explicitly listing AI-generated sexualized impersonation of real people as a contextual signal of non-consensual intimate imagery to address the harms of such content.

Access to Remedy

65. Similar to Explicit AI Images of Female Public Figures, in the present case, the initial reports and appeals of Meta's decisions were not prioritized for human review on time. In Explicit AI Images of Female Public Figures, the Board noted that regardless of the public or non-public figure status of victims, a delay in removing these images severely undermines their privacy and can be catastrophic. Echoing the discussion and the recommendation in that case on improving reporting flows on AI-generated non-consensual intimate imagery, the Board reiterates that access to remedy in these contexts is an issue that may have significant human rights impacts that require careful risk assessment and mitigation.

6. The Oversight Board's Decision

The Board overturns Meta's decision to leave up the content and requires that the post be removed.

7. Recommendations

Policy

1. To 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 sexualized impersonation of real people.

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

Enforcement

2. To reduce the burden on victims of non-consensual intimate image abuse, Meta should allow users to designate connected accounts, such as those of trusted friends, family members or associates, that can report on their behalf potential Community Standards violations involving non-consensual intimate imagery, including impersonations.

The Board will consider this recommendation implemented when Meta makes these features available and easily accessible to all users via their account settings.

3. To ensure effective and accessible reporting pathways to report on AI-generated non-consensual intimate imagery abuse, Meta should include AI-generated sexualized impersonation as a separate category in standard content reporting and appeal forms, distinct from "harassment" or "nudity." These uniform reporting forms should be available globally.

   The Board will consider this recommendation implemented when Meta updates its reporting forms, shares data showing the global rollout results and explains how the new reporting flows impact enforcement of the Adult Sexual Exploitation Community Standard.

The Board also reiterates the importance of its previous recommendations, noting their relevance to the issues of this case. In line with those recommendations, Meta should:

- Change the word "derogatory" in the prohibition on "derogatory sexualized photoshop" to "non-consensual" (Explicit AI Images of Female Public Figures, recommendation no. 2).
- Replace the word "photoshop" in the prohibition on "derogatory sexualized photoshop" with a more generalized term for manipulated media (Explicit AI Images of Female Public Figures, recommendation no. 3).
- Implement content credentials (as laid out by the C2PA) at scale and ensure that they are clearly and consistently visible and accessible to users whenever the provenance details are available (AI-Generated Video in Israel-Iran Conflict, recommendation no. 4).

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.