Board to Address Non-Consensual AI Sexualized Impersonation

Digital Rights Foundation
26 February 2026

The Oversight Board's February 2026 case selection concerns a reported AI-generated, non-consensual sexualized video impersonating a woman on Instagram, where multiple user reports were not prioritized for human review. Meta ultimately decided the post did not violate its Adult Nudity and Sexual Activity rules and applied an "adults only" restriction, reasoning that the video focuses on the subject's crotch/underwear area.

Nonconsensual AI-generated images and videos are disturbingly prevalent across platforms. Recently, Elon Musk's AI chatbot Grok came under fire for this exact reason. The lax regulation of prompts pertaining to image generation on Grok led to the pictures of hundreds of thousands of women X users being misused. Bad actors prompt AI chatbots such as Grok to remove articles of clothing of existing pictures of women, depict them in lingerie/swimsuits[1]/torn clothing[2], or covered in semen[3] or semen-like substances[4].

In South Asia, a similar prevalence is seen on platforms such as Instagram, Facebook, TikTok, and X. As early as 2023, there was an upsurge in nonconsensual explicit deepfakes of Bollywood female actors and other public figures, one case[5] out of which was also taken up by the Oversight Board in April 2024. Similarly, in Pakistan, deepfakes of popular Pakistani actors kissing also circulated in Pakistan. Though ostensibly made with non-malicious intent, they were still deeply harmful to the women actors, as they had expressed[6]. Moreover, there have been videos of women politicians circulating on platforms, a lewd deepfake video targeted[7] Azma Bokhari, the information minister of Punjab, in 2024. Apart from high-profile women, regular women and girls in Pakistan are routinely targeted using AI-manipulated and doctored images and videos, with doctored vulgar slogans[8] being attributed to the annual Women's Day Aurat March protestors being a prime example. In fact, the Digital Rights Foundation's Digital Security

---

[1] https://www.reuters.com/legal/litigation/grok-says-safeguard-lapses-led-images-minors-minimal-clothing-x-2026-01-02/

[2] https://x.com/grok/status/2023033015044030805?s=20

[3] https://www.theguardian.com/technology/2026/jan/08/grok-x-nonconsensual-images

[4] https://www.theguardian.com/technology/2026/jan/05/elon-musk-grok-ai-digitally-undress-images-of-women-children

[5] https://www.oversightboard.com/news/oversight-board-announces-two-new-cases-on-explicit-ai-images-of-female-public-figures/

[6] According to a caller at the Digital Rights Foundation's Digital Security Helpline, one actor's brother said explicitly that they knew the pictures were fake, but if anyone else saw them he would kill her. In such a risky situation where a person's dignity and physical safety is in danger, any arguments of artistic expression, freedom of expression, and age gating lose much of their relevance.

[7] https://www.dawn.com/news/1848313

[8] https://digitalrightsfoundation.pk/gendered-disinformation-the-war-on-women-and-gender-minorities/

Helpline began receiving cases of GenAI non-consensual image-based abuse only after they were used exponentially during the 2024 Pakistani elections to target women journalists, politicians, and other public figures.

Meta has previously introduced region/country/escalation-specific policies and enforcements. However, this is not one of those areas where those conditions should be applied. GenAI NCII impacts women all over the world. Women and girls from certain countries might face more dire consequences, but people from such cultures or regions might reside in other 'global North' countries, which might be perceived as safer for women, as well. The DRF Helpline has had multiple cases where Pakistani women residing in Italy or the UK have been killed for honor either there or after being forced to come to Pakistan.

The issue is not limited to one platform, nor does it only target women. According to a UNICEF brief[9] from February 2026, the UK's Internet Watch Foundation found nearly 14,000 suspected AI-generated child abuse images on a dark web forum, with about a third confirmed as illegal, along with the first realistic AI-generated abuse videos. Large-scale research across 11 countries has shown[10] that around 1.2 million children reported images manipulated into sexually explicit AI deepfakes this past year alone.

Women and children who use social media platforms have the right to post their images and videos on these platforms without the fear of having someone manipulate or morph them into sexualised content using AI tools. This, besides being a point raised by the presiding judge in a case[11] where the perpetrator took images of women from social media pages and created deepfake porn using them, is the guiding principle upon which platforms need to structure content moderation policies.

The core foundation of all policies should be looked into to set an appropriate context for cases under review. The Meta Adult Nudity and Sexual Activity Community Standard was designed to protect the freedom of expression of women who chose to express themselves in a certain way, but when considering non-consensual cases, the Bullying and Harassment Community Standard should be investigated alongside. This policy focuses on private individuals: "For private individuals, our protection goes further: We remove content that's meant to degrade or shame, including, for example, claims about someone's sexual activity...Context and intent matter…"[12] This core foundation is more applicable here, and creating sexualized images of women and girls, especially ones that reveal underwear, can be understood to degrade or shame. "Content sexualizing another adult" falls under Tier 2 of the bullying and harassment policy, but needs to be better enforced with greater restrictions to avoid cases like the one under review.

In this vein, the protection of the image/likeness of users on social media platforms must take precedence over artistic expression in cases where images of likenesses of real women are being manipulated or warped through AI to be sexually repurposed. Moreover, the "artistic expression" justification can easily

[9] https://www.unicef.org/media/178571/file/UNICEF%20AI%20CSEA%20Brief_2.pdf
[10] https://www.thorn.org/press-releases/report-1-in-10-minors-say-peers-have-used-ai-to-generate-nudes-of-other-kids/
[11] https://www.bbc.com/news/articles/cewgxd5yewjo
[12] https://transparency.meta.com/policies/community-standards/bully ing-harassment/

be misused by anyone with malicious intent, too. Protection should not be limited to faces and bodies, but voices as well. A disturbing recent trend reveals dozens of AI-generated videos[13] taking as their blueprint the video of a woman singing on Instagram. Her voice is nonconsensually edited onto several AI-generated depictions of women spanning various races, with each viral AI video racking up views in the millions. Platforms need to start proactively detecting and removing such non-consensual AI images and videos, especially those of a sexual or sexually suggestive nature. At the very least, notice-and-takedown policies need to be strictly enforced, with such cases being prioritised for human review, which Meta failed to do in the instant case.

From a Global South civil society perspective, the case sits at the intersection of automation-driven enforcement gaps and the "offline consequences" of sexualized image misuse, including harassment, reputational harm, extortion, employment impacts, and physical safety risks, especially in contexts where gender norms, stigma, and weak institutional remedies intensify harm.

The Oversight Board has previously noted how socially conservative environments can amplify harms from sexual deepfakes and other manipulated sexual imagery, including serious real-world violence triggered by altered images. A key lesson from this case's fact pattern is that "age restriction" and "no policy violation" outcomes can become a de facto dismissal of the core allegation around non-consensual sexualized impersonation.

For civil society organizations supporting survivors in low-resource environments, this can be experienced as a denial of remedy because reputational damage and community enforcement (shaming, threats, coercion) continue regardless of whether viewers are under 18 or over 18.

Non-consensual AI sexualized impersonation should be treated as non-consensual sexual content, not merely as a question of adult nudity. The core harm lies in the lack of consent and reputational damage, not in whether the content meets a pornography threshold. The Oversight Board has previously emphasized that deepfake intimate imagery disproportionately harms women and girls, and that removal, rather than labeling or age restriction, is the only effective remedy.[14] A rights-based approach would therefore explicitly prohibit AI-generated or AI-manipulated sexualized depictions of real individuals without full consent, including suggestive (not just explicit) content.

Platforms should also adopt a "presume non-consent" standard for private individuals when credible signals of impersonation or AI manipulation exist, rather than requiring public visibility or media coverage as proof.

---

[13] https://www.instagram.com/p/DUX_gObjBqG/?utm_source=ig_web_button_share_sheet

[14] https://www.oversightboard.com/news/new-decision-addresses-metas-rules-on-non-consensual-deepfake-intimate-images/

Age-gating alone is not an adequate response. Restricting content to adults does not prevent reputational harm, harassment, re-uploads, or cross-platform spread, and weak age verification systems are easily bypassed. At the same time, stricter age verification can raise privacy and equity concerns, particularly in Global South contexts. Instead, platforms should treat such cases as high severity and ensure rapid human review, prevent auto-closure of reports, suspend distribution pending review, and use hashing and media-matching tools to block re-uploads.

Dedicated reporting pathways for "AI sexualized impersonation," privacy-centered reporting mechanisms, and priority escalation channels for trusted partner civil society networks are essential to ensure effective, survivor-centered enforcement.