WITNESS Public Comment on Non-Consensual AI Sexualized Impersonation Case

1. Introduction

WITNESS submits this public comment as a human rights organisation with over thirty years of experience in media technology, documentation, and accountability, with particular expertise in AI governance, platform accountability, and how AI-generated content affects evidence verification and gender-based violence.

In July 2024, WITNESS submitted public comment to the Board's Explicit AI Images of Female Public Figures case. The Board cited WITNESS twice in that decision, and our analysis shaped the Board's consent-centred framing of non-consensual intimate imagery (NCII) and its critical separation of NCII harms from the broader "manipulated media" detection frame.

The Board made four recommendations. To date, Meta has not meaningfully implemented them.

Meta's September 2024 response to the four recommendations from the Board declined the most important (moving the prohibition into Adult Sexual Exploitation), deferred the most urgent (treating AI-generated context as a signal of non-consent), and agreed in principle only to updating "photoshop" terminology while ruling out replacing "derogatory" with "non-consensual." Meta's own Transparency Center page states that it "do[es] not expect to replace 'derogatory' with 'non-consensual'" and "do[es] not expect that this will result in moving the prohibition." These are not pending implementations. They are documented refusals. The Bullying and Harassment policy still contains the phrase "derogatory sexualized photoshop or drawings," and Meta told CBS News in February 2025 that it was "still considering" how to treat AI-generated context as a non-consent signal, seven months after the Board's recommendation. In the case now before the Board, Meta's subject matter experts assessed the content under Adult Nudity and Sexual Activity, a policy with no framework for consent, rather than the Adult Sexual Exploitation standard the Board recommended in 2024, or even the Bullying and Harassment standard where Meta itself insists the prohibition should remain.

Meanwhile, a June 2025 CBS News investigation found hundreds of nudification app advertisements running across Meta's platforms, removed only after publication.

Every vulnerability the Board identified in 2024 including over-reliance on media reports as a consent proxy, auto-closure without human review, policy routing that diverts NCII from appropriate enforcement, recurs here. But the victim is a non-public figure: ***the exact population the Board warned was most at risk.***

Since PC-27095, WITNESS has continued to build its evidence base: our March 2025 submission to the UN Human Rights Council documented AI-driven sexual violence as a top-tier threat; our December 2025 submission to the Board's Iran-Israel case addressed likeness protection; and our analysis of the January 2026 Grok NCII cases demonstrated that harm occurs at creation, not only distribution.

This case highlights a structural failure to treat AI-generated sexualized impersonation as sexual exploitation.

2. Tech facilitated Gender based violence

AI generated NCII is a form of technology-facilitated gender-based violence that disproportionately affects women. A recent study by the Oxford Internet Institute analyzed nearly 35,000 deepfake models available for public download and found out that 96% of these models targeted identifiable women. Platforms must address this as image-based sexual abuse, not merely as a content moderation challenge.

3. Responses to the Board's Questions

Question 1: Prevalence and Use of AI-Generated NCII

AI-generated NCII has rapidly shifted from fringe abuse forums into scalable, commercialized and mainstream distribution ecosystems. The key development is not only increased volume, but increased ease, plausibility, and reach. The issue is no longer episodic misuse. It reflects structural features of generative systems, advertising infrastructure, and enforcement design. We highlight six interrelated trends.

1. **Industrialization of abuse:** Consumer-facing nudification and deepfake tools have dramatically lowered the barrier to producing sexualized impersonations at scale. What distinguishes the current moment is the embedding of these tools into commercial ecosystems: nudification apps advertised on mainstream platforms, abuse services operating through subscription models, and platform advertising systems functioning as discovery infrastructure for exploitation.
2. **Gendered targeting:** As documented in WITNESS' March 2025 submission to the UN Human Rights Council, AI-driven sexualized impersonation disproportionately targets women, girls, LGBTQ+ individuals, journalists, and human rights defenders. It functions as technology-facilitated gender-based violence (TFGBV), reinforcing patterns of misogyny, coercion, humiliation, and silencing. For non-public figures, consequences including reputational harm, employment loss, extortion, psychological trauma, frequently occur long before content moderation concludes.
3. **Likeness exploitation and the erosion of consent signals:** High-fidelity generative systems (including tools such as OpenAI's Sora) mark a structural shift in the synthetic media ecosystem, enabling increasingly precise exploitation of likeness without consent. As realism increases, harm expands beyond explicit nudity to include sexualized impersonation that may not meet strict nudity thresholds but nonetheless constitutes intimate exploitation. Platforms often rely on contextual cues of non-consent such as media reporting or public figure status, yet AI-generated impersonation frequently targets non-public figures in private networks where no public documentation exists. The absence of media coverage cannot be treated as evidence of consent.
4. **Plausible deniability and verification collapse:** Increasing realism enables perpetrators to evade accountability through plausible deniability ("it's fake," "it's a joke"), shifting the burden onto victims.

Technical safeguards such as watermarking remain inconsistent and are often stripped across platforms. And as WITNESS documented, detection tools are unreliable in real-world conditions, particularly across multimodality, file quality, and global contexts. This is compounded by platforms' own failure to identify synthetic content on their services: an Indicator audit found that platforms fail to label AI-generated content 70% of the time, including content produced with their own tools. Enforcement frameworks that depend on proving AI generation or conclusively verifying manipulation will therefore systematically fail victims.

5. **Harm at creation:** The January 2026 Grok incident illustrates that harm occurs at the moment of synthetic creation when a system generates non-consensual sexualized depictions upon prompt, the violation has already occurred regardless of distribution. AI-generated NCII is not only a moderation issue; it is a product design and system safeguards issue.
6. **Capture-to-exploitation pipeline:** The proliferation of AI-enabled wearable devices is creating new vectors for non-consensual sexualized impersonation. Smart glasses equipped with discreet cameras, including Meta's Ray-Ban smart glasses, enable covert capture of likeness in public and semi-public spaces. In February 2026, a Russian national used such devices to secretly record encounters with women across Kenya and Ghana, monetising the footage through a multi-platform pipeline: free clips on TikTok and YouTube driving traffic to a paid Telegram channel, the same advertising-to-exploitation model documented in the nudification app ecosystem Kenya's Ministry of Gender, Culture and Children Services characterised the incident as "a serious form of technology-facilitated gender-based violence and exploitation." As global smart glasses shipments grew 210% in 2024 and major manufacturers integrate generative AI capabilities directly into wearable hardware, covertly recorded imagery can increasingly be fed into nudification or synthetic impersonation tools with minimal friction, creating a seamless pipeline from capture to abuse. This is particularly relevant to the Board because Meta is simultaneously the platform where NCII circulates and quickly becoming a manufacturer of the hardware that facilitates covert likeness capture.

Question 2: Likeness Protection and Artistic Expression

The core principle is consent. There are categorical distinctions between: transformative or satirical expression that does not sexualise an identifiable individual; sexualized impersonation that exploits a person's likeness without consent; and consensual use of synthetic media by individuals, including sex workers, to protect their identity or exercise agency over their own likeness. The distinguishing principle across all three is consent.

The recent Kenya and Ghana incident described above underscores this. As Kenyan journalist Ferdinand Omondi wrote: "Consent to sex is not consent to filming. Consent to filming is not consent to publication." Neither artistic nor commercial framing can cure non-consensual exploitation.

International human rights law protects expression, but also dignity, privacy, and bodily autonomy. Consent must be the governing standard. In practice: platforms should treat realistic AI-generated sexual content depicting identifiable individuals as presumptively non-consensual; public figure status does not nullify consent requirements; artistic exemptions should not apply where sexualisation is involved; and

likeness protection should not depend on whether content is labelled as AI-generated. The violation lies in the non-consensual sexualisation, not the technical method.

Question 3: Enforcement Best Practices

a. Correct policy architecture: The prohibition on non-consensual sexualized manipulated media should be moved from Bullying and Harassment into the Adult Sexual Exploitation (ASE) standard, where enforcement expertise is aligned with exploitation harms. The requirement that content be "non-commercial or produced in a private setting" should be removed: AI-generated content is by definition not produced in a private setting, and the nudification ecosystem is inherently commercial. Retaining this language creates a loophole excluding the precise harm this case involves. AI-generated context should be treated as a signal of non-consent — consent to share an image publicly does not mean consent to its sexualised manipulation.

b. Adopt a presumption of non-consent: For realistic AI-generated sexual content depicting identifiable individuals, platforms should presume non-consent unless affirmative, verifiable consent is demonstrated. Public figure status and absence of media coverage cannot serve as proxies for consent. As Danielle Citron has argued, public revelation of a person's sexual expression without consent interferes with autonomy and self-respect. The burden should not rest on victims.

c. Mandate human review and prohibit auto-closure: Reports involving AI-generated sexual impersonation must not be auto-closed. "Intimate" content is highly contextual and cannot be reliably inferred from nudity alone — as this case demonstrates. There is currently no independent benchmark for evaluating NCII detection across diverse content types, file qualities, and cultural contexts. WITNESS sees a need to extend its TRIED AI Detection Benchmark to NCII. Until such evaluation exists, automated detection claims cannot substitute for human review.

d. Address creation-level risk through product safeguards: Platforms deploying or integrating generative systems should implement: prompt-level safeguards preventing sexualized impersonation of identifiable individuals; likeness-protection guardrails for everyone; escalation protocols when models generate prohibited impersonations; and responsible design standards for AI-enabled capture devices (including smart glasses) to prevent covert likeness capture from feeding into exploitation pipelines, particularly where, as with Meta, the hardware manufacturer and the content platform are the same company.

e. Transparency as a precondition: Without independently verifiable transparency, enforcement claims lack credibility. Building on WITNESS's 2024 submission and CDT's Model NCII Policy, platforms should publish detection rates, false positive/negative rates, removal timelines, geographic and language disparities, and appeals outcomes.

Question 4: Effectiveness of Age-Gating

Age-gating is not an effective safeguard against AI-generated NCII. Age verification mechanisms can be easily circumvented, and the core harm is tied to consent and impersonation, not viewer age. Age-gating does not prevent the creation of abusive content and does not meaningfully address the exploitation of victims, many of whom are adults.

Age-based restrictions should complement enforcement, but they cannot substitute for consent-based prohibition and rapid removal mechanisms.

Question 5: Reporting Mechanisms

Effective reporting requires: a dedicated category for AI-generated sexualized impersonation distinct from harassment or nudity; evidence-sensitive design allowing victims to report without re-uploading abusive content; removal of requirements to demonstrate public visibility or private-setting production; time-bound specialist review rather than auto-closure; and cross-platform coordination through mechanisms such as StopNCII, with due process safeguards.

3. Recommendations

WITNESS agrees with and urges the Board to reiterate its 2024 recommendations, which remain unimplemented. We consider the need to change the framing of Synthetic NCII, lack of consent in the generation of these contents and call to strengthen transparency still timely and needed in regards to many of the points highlighted in our submission.

Besides that, we would like to recommend the following:

1. **Recognize AI-generated NCII as technology-facilitated gender-based violence**, aligning policy language accordingly.
2. **Adopt a presumption of non-consent** for realistic AI-generated sexualized impersonations of identifiable individuals, regardless of the fact that such images might have been generated based on media made available by the individuals themselves - as it implies a manipulation of the purposes regarding the image uses.
3. **Mandate human review** for all AI sexual impersonation reports.
4. **Create a dedicated reporting pathway** for AI-generated sexualized impersonation.
5. **Implement product-level safeguards** to prevent the generation of sexualized impersonations at the system design stage.
6. **Extend product accountability to AI-enabled capture devices.** Meta should establish responsible design standards for its hardware products to prevent covert likeness capture from feeding into exploitation pipelines.