Comment on Meta's Oversight Board on Non-Consensual AI Sexualized Impersonation

Authors: The Integrity Institute[1] — Theodora Skeadas, Sarah Amos, Leah Ferentinos, Gabe Freeman, Rupi Sureshkumar, Sofia Bonilla, Matt Motyl, Aniket Ajagaonkar, April Lat

SUMMARY

First, Meta assessed this content under the wrong policy frame. The video was explicitly reported as a non-consensual AI-generated impersonation, yet Meta evaluated it solely under its Adult Nudity and Sexual Activity standard — concluding it did not meet the threshold for removal and applying an age restriction instead. This case involves two distinct harms, each requiring its own policy response. The first is non-consensual intimate imagery: content that depicts a person in a sexualized or intimate manner they did not consent to. The second is non-consensual impersonation: the fabrication of a person's likeness without their permission — a violation of autonomy that exists independent of whether the content is sexual. Meta's general sexual-content lens cannot assess either form of non-consent. And age-gating is the wrong remedy for a consent violation: the harm to the depicted individual persists regardless of the viewer's age.

Second, Meta needs both stronger policy and a functioning enforcement pipeline. On the policy side, Meta has no framework for "someone fabricated my likeness without consent" as a distinct harm, and its definition of "intimate" is too narrow — anchored to visible nudity rather than the broader violations of dignity that generative AI now makes trivially easy to produce. On the pipeline side, this content was reported multiple times by multiple users as non-consensual AI-generated impersonation. One user appealed. No report ever reached human review. Because both intimate-imagery and impersonation harms are inherently contextual — they cannot be resolved by automated classifiers alone — the reporting-to-review pipeline is where the system must work. In this case, it failed entirely.

Third, Meta bears heightened responsibility as both an AI developer and a high-reach distribution platform. Through Llama, Meta releases open-source models whose safety guardrails can be stripped via fine-tuning or jailbreak. Through Instagram and Facebook, it operates the platforms where this content is distributed at scale. This dual role demands a correspondingly higher standard of responsibility — particularly at the distribution layer, where Meta has direct enforcement capability.

We recommend that Meta:

1. Restructure its enforcement pipeline so that content reported as non-consensual intimate imagery or non-consensual impersonation is routed to dedicated human-review queues with rapid response targets, rather than evaluated solely as adult content.
2. Implement prevention-layer interventions — including trust-gated posting, upload-time nudges, and proactive detection — that reduce the volume of non-consensual intimate content reaching audiences in the first place.
3. Redesign its reporting flows to be globally accessible, discoverable, and responsive, with support for third-party reporters and connections to survivor resources.

Q1: PREVALENCE AND CONTEXT

The scale of AI-generated non-consensual intimate imagery and video is well-documented and accelerating. Other organizations submitting comments will address prevalence and survivor impact in greater depth; we focus here on what the Integrity Institute's membership is positioned to assess: the analytical framework needed to address this problem and the operational dynamics that make it particularly resistant to existing platform enforcement. Two points underpin the rest of our comment.

---

[1] The Integrity Institute submits this comment drawing on the expertise of our membership, which includes current and former Trust & Safety professionals who have designed and operated content moderation and enforcement systems at major platforms. Several members contributed as co-authors to three resources this comment draws on substantially: *Digital Violence, Real World Harm* (Humane Intelligence & UK Government), *Prevention by Design: A Roadmap for Tackling TFGBV at the Source* (Search for Common Ground, Integrity Institute & Council on Technology and Social Cohesion), and the *TFGBV Taxonomy* (Humane Intelligence).

First, "intimate" must be understood broadly. Meta's current definition is anchored to visible nudity and sexual encounters — calibrated for a pre-generative-AI world in which intimate imagery required an original intimate moment. That is no longer the case. AI-generated content can fabricate violations of dignity that bear no relationship to nudity or sexual acts — for example, depicting a person who normally wears hijab without it. Meta's narrow definition caused it to miss the harm in this case entirely.

Second, generative AI has fundamentally changed the structural conditions under which both harms operate. Open-source models can be downloaded, run locally, and fine-tuned with safety guardrails stripped entirely, as documented in *Digital Violence, Real World Harm* (pp. 11, 26). The January 2026 Grok incident demonstrated that even commercial AI platforms may launch with insufficient protections. Advances in generative AI have lowered the barrier for non-technical actors to produce convincing fabricated images and video at a frequency and scale that was not possible even two years ago. Because not all AI systems will prevent this content at the generation layer, the burden necessarily falls on distribution platforms to detect and act on it.

Q2: PROTECTION OF IMAGE AND LIKENESS

The non-consensual use of a person's likeness is a harm independent of whether the resulting content is sexual. Individuals must be able to meaningfully control the generation and dissemination of AI-created images and videos depicting their likeness — both to prevent intimate image abuse and to prevent impersonation. This principle applies across two domains: AI development (where content is generated) and content distribution (where it reaches audiences). Meta's failure in this case was not only an enforcement gap but a conceptual one: the company has no policy framework for "someone fabricated my likeness without consent" as distinct from "this content is sexually explicit."

At the generation layer, AI image- and video-generation systems should not permit the creation of content depicting a real person's likeness without that person's explicit consent — and this is especially urgent in sexual or intimate contexts. OpenAI's Sora offers a useful reference point: its "cameo" feature requires individuals to record themselves and actively opt in before their likeness can be used, with granular controls over who may generate content featuring them and the ability to revoke access at any time. The Sora model is imperfect — enforcement gaps emerged quickly after launch — but the underlying design principle is sound: consent to the use of one's likeness should be affirmative, specific, and revocable.

Artistic expression remains fully preserved under this framework. Creators can generate AI content depicting fictional individuals without restriction. The line is drawn at using a real person's likeness without their consent — whether that person is a public figure or not. We acknowledge this boundary involves edge cases — an AI-generated figure may closely resemble a real person unintentionally — but the existence of hard cases does not justify the absence of a clear default rule.

At the distribution layer, platforms should detect and act on non-consensual use of a person's likeness, with particular urgency for intimate or sexualized content, and offer accessible, rapid removal mechanisms.

Q3: BEST PRACTICES AND RECOMMENDATIONS FOR ENFORCEMENT

Prevention

The most effective interventions stop harm before content reaches an audience. The prevention mechanisms below primarily address the sexually explicit content dimension — they rely on mature content classifiers that can identify such imagery at upload. The impersonation dimension is inherently harder to catch proactively, because it requires knowing whose likeness is depicted and whether they consented. This makes the enforcement and reporting layers (addressed below and in Question 5) critical: impersonation harms are primarily surfaced through user reports, and the system fails entirely if those reports are not routed to review — which is precisely what happened in this case. Three prevention mechanisms are deployable with existing platform capabilities:

- **Rate limits on low-trust accounts.** Platforms should impose rate limits and feature restrictions on

low-trust accounts — newly created, unverified, or with limited interaction history — before permitting them to post sexually suggestive content. This is standard practice for other high-risk features such as livestreaming and monetization, and should extend to content categories with high abuse potential.

- **Upload-time nudges.** When content is flagged at upload as potentially explicit or intimate by existing media classifiers, platforms should prompt users to reconsider before posting. This intervention has strong evidence behind it: Instagram reported that when it sent approximately one million nudges over the course of a single week, users deleted or amended their content 50% of the time (Prevention by Design, p. 5). Nudges also create a stronger remediation basis — an account that posts after being warned has demonstrated intent, which simplifies downstream enforcement decisions.

- **Proactive monitoring.** Platforms should monitor for distribution spikes of sexualized content featuring the same individual — a signal of coordinated distribution or viral non-consensual content. We note that while some organizations advocate for AI-generated content detection at upload, current deepfake detection classifiers are not reliable enough to operate at Meta's scale with acceptable error rates, and watermarking only covers compliant commercial models. The more operationally sound approach is to focus detection on the content characteristics that make it harmful — intimate imagery, non-consensual distribution patterns — rather than on whether it was AI-generated. Meta already possesses these classifiers; this case failed not because detection capability was absent, but because reports were not routed to human review.

Enforcement

- **Repeat offender controls.** Accounts confirmed to have posted NCII or non-consensual impersonation content should face escalating sanctions: account suspension, device-level restrictions, and contribution of abuse signals to cross-platform databases.

- **Victim-centric protections.** Two distinct mechanisms address two distinct threats. For known NCII, Meta already participates in hash-matching through StopNCII.org — a foundation that should be maintained and deepened. The gap is in novel AI-generated impersonation content, where no reference image exists to match against and the victim cannot predict what the content will look like. A **likeness consent flow** — in which individuals can register their facial features to flag unauthorized use of their likeness — offers a more scalable path for this category of harm. However, we flag a real tension: asking victims to opt into facial recognition to protect themselves raises concerns about bias, surveillance, and disproportionate impact on marginalized communities with documented higher false-positive rates in facial recognition systems. Any such system must be strictly opt-in, purpose-limited exclusively to abuse prevention, and subject to independent bias auditing. Meta, which already operates facial recognition infrastructure, is positioned to pilot this — but only with these safeguards in place.

- **Closed groups.** Meta should proactively investigate closed groups that share links to known nudifying applications or distribute intimate imagery at scale — these groups function as infrastructure for producing and distributing non-consensual content, and addressing them disrupts harm upstream of individual posts.

Consent Determination

Meta currently determines non-consent based on three signals: vengeful context in captions or comments, independent sources such as law enforcement records, and a report from the person depicted (Transparency Center). This framework is too narrow, and this case demonstrates why: a friend of the depicted person reported the content as non-consensual AI-generated impersonation, and the report was not prioritized. Three changes are needed:

- First, **third-party reports** alleging non-consent should carry material weight in review prioritization. Third-party reporters may be the first to discover the content — but they may also be trusted individuals to whom the victim has delegated the administrative burden of navigating platform reporting processes. Both roles matter, and the system should accommodate both.

- Second, when a report alleges that content depicts a real person without consent, **non-consent should be presumed** and the burden should rest on the poster to demonstrate that the depicted individual's

consent was obtained — for example, through a verifiable opt-in record, identity confirmation that the poster is the person depicted, or documentation of a consent agreement. This is especially critical for AI-generated content, where the imagery is entirely fabricated and no original moment exists at which consent could have been given. The current framework, which effectively requires the victim to prove non-consent after the fact, inverts the burden for the party least equipped to bear it.

- Third, **consent is revocable** — even imagery originally shared consensually can become non-consensual, and enforcement systems must account for this.

Q4: AGE-GATING

Non-consensual intimate imagery — whether AI-generated or not — should be removed, not age-gated. Age-gating, Meta's response to the appeal in the case in question, is the wrong remedy because it applies a content-sensitivity framework to what is fundamentally a consent violation. The harm to the depicted individual is not that the content is sexually suggestive; it is that their likeness was used without permission to create a sexualized impersonation. That harm persists regardless of the viewer's age. This case illustrates the failure directly: Meta age-restricted the video rather than removing it, leaving the content accessible to adult users while the depicted person's dignity and autonomy remained violated. Age-gating may be appropriate for lawful adult content; it is not appropriate for non-consensual content, which should not exist on the platform at all.

Q5: REPORTING MECHANISMS

This case is not evidence of a policy gap alone — it is evidence of a pipeline failure. The content was reported multiple times by multiple users as non-consensual AI-generated impersonation. One user appealed. At no point did any report reach human review. Meta's automated systems assessed each report algorithmically and got every determination wrong. The lesson is not that Meta needs a better policy; it is that Meta needs a reporting pipeline that routes these reports to someone who can actually evaluate them.

The Report-to-Review Pipeline

The most direct lesson of this case is architectural: reports alleging non-consensual intimate imagery or non-consensual impersonation cannot be resolved algorithmically. Consent and likeness questions are inherently context-dependent. Automated systems can triage and prioritize, but final determinations on these report categories must involve human judgment. Meta must build **human review as a hard requirement** into its pipeline for both NCII and impersonation reports — not as an aspiration, but as a system constraint. Two operational changes follow directly:

- **Dedicated fast-track queues.** NCII and non-consensual impersonation reports should enter a queue separate from general content reports, with a target of human review within hours. Surveys of NCII survivors consistently report response times of days to weeks, with many reports receiving no response at all (p. 11, 44). NCII should already receive expedited treatment in theory; survivor experience indicates it does not in practice.

- **Suppress algorithmic amplification pending review.** While a report is awaiting review, the content should be removed from recommendations, made ineligible for resharing, and stripped of monetization. Algorithmic amplification is a platform privilege, not a default entitlement. The cost of temporarily withholding amplification from content that turns out to be legitimate is far lower than the cost of algorithmically distributing a non-consensual intimate video or fabricated likeness while a review queue clears — which is what happened in this case.

Gaps in Meta's Current Reporting Flows

Meta's reporting infrastructure has several gaps that this case exposes. We note that Meta has taken meaningful steps — including a dedicated NCII reporting form that distinguishes between intimate imagery and deepfake intimate imagery, a response to civil society advocacy for a direct, accessible path that is not buried in nested menus. The gaps below are not about the existence of that path but about what happens around and beyond it.

- **No reporting path for non-consensual impersonation as a distinct harm.** The Board specifically noted that this case involves impersonation, not just sexual content. Meta's reporting taxonomy does not reflect that distinction. There is no clear path to report "someone fabricated my likeness without consent" as a standalone harm — reports about AI-generated impersonation are forced into categories designed for sexual content or general intellectual property, neither of which captures the harm. Adding a dedicated impersonation path would ensure these reports trigger the right review.

- **Global availability.** Meta should ensure its dedicated NCII reporting path is accessible globally — not limited to jurisdictions that have passed specific NCII legislation.

- **Narrow framing of support resources.** Meta's current NCII support page is structured around sextortion, despite the 2019 announcement describing a broader scope of harm. NCII encompasses AI-generated impersonation, non-consensual nudification, and distribution without threat or coercion. This narrow framing encodes the same assumption that caused the enforcement failure: that non-consensual intimate imagery requires a threat or coercive context to be harmful. It does not. The support experience should be realigned to cover the full range of NCII and non-consensual impersonation.

Design Recommendations

- **Quarantine system for gray-area content.** When a high-trust account flags content from a low-trust or newly created account, the platform should remove the content by default pending review. In the reverse scenario, the content enters the review queue but is not automatically removed. This approach, detailed in the Integrity Institute's Prevention by Design report (p. 13), is the enforcement-side complement to the rate limits on low-trust accounts we recommend in Question 3: new and unverified accounts should face both higher friction to post potentially harmful content and faster removal when that content is flagged.

- **Trusted third-party reporters.** Victims of NCII are often unaware the content exists, may lack secure access to a personal device, or may not have the digital literacy to navigate reporting flows. This case itself was reported by a friend, not the depicted individual. Meta should allow users to designate trusted individuals — friends, family members, or NGO caseworkers — who can report on their behalf. Meta already has an analogous model: the legacy contact system for deceased users' accounts. More broadly, Meta should continue and deepen its investment in local NGOs as trusted flaggers. As the scale of AI-generated intimate imagery increases, these organizations' caseloads will grow correspondingly, and platform investment must keep pace.

- **Prioritize NCII and impersonation reports in review queues.** Non-consensual intimate imagery is a time-sensitive harm: every hour content remains accessible compounds the damage. NCII and non-consensual impersonation reports should be triaged ahead of less urgent content categories, with defined response-time targets comparable to those applied to child sexual abuse material.

Cross-Platform and Systemic Recommendations

- **Hash-sharing through StopNCII.** Meta already participates in StopNCII.org and has publicly committed to sharing hashes of NCII it removes from its own apps with other platforms through the StopNCII infrastructure. The Board should recommend a concrete timeline for fulfilling this commitment — particularly for AI-generated content flagged through Meta's reporting flows, which represents the fastest-growing category of harm.

- **Interoperable evidence documentation.** Separately from hash-matching, survivors currently lack standardized tools to collect and preserve evidence of abuse — screenshots, metadata, URLs — in formats that are usable across platforms and in legal proceedings. Meta should support the development of interoperable documentation tools, potentially in partnership with civil society organizations already doing this work (such as Pirth.org), to reduce the burden on survivors who must currently navigate each platform's reporting process independently.

- **Industry-wide response-time standards.** Meta should work toward an industry norm — modeled on the urgency applied to CSAM — that commits participating platforms to defined response windows for confirmed or high-confidence NCII and non-consensual impersonation reports.