CEE Digital Democracy Watch

Overseeing the New Digital Era in Election Integrity

Warsaw, 26.02.2026

Public Comment — Case 2026-021-IG-RA
Reported AI-Generated Sexualized Video
Submitted by: CEE Digital Democracy Watch

This comment is submitted in response to Case 2026-021-IG-RA, in which the Oversight Board examines an Instagram video allegedly generated using artificial intelligence and depicting a woman in a sexualized manner without her consent, as well as Meta's response to and handling of the incident.

The case touches on a critical challenge in the contemporary information ecosystem: the use of generative AI to create non-consensual, sexualizing representations of real individuals. In the literature, such material is variously referred to as deepfake pornography, non-consensual intimate imagery (NCII), or non-consensual sexualizing imagery (NSII). CEE Digital Democracy Watch advocates for the term non-consensual sexualizing deepfakes as most accurately capturing the abusive and harmful nature of the phenomenon. For the purposes of this comment, we use the acronym NSII.

Our analysis draws on the findings of our 2025 report, *Non-consensual Sexualising Deepfakes: Threats, and Recommendations for Legal and Societal Action*, which provides an empirical basis for the concerns and recommendations set out below.

CEE Digital Democracy Watch
Mokotowska 43/104, 00-551 Warsaw, Poland
www.ceeddw.org
KRS 0001090110 | EU Transparency Register 114284692189-05

Introduction

Meta's decision to retain the content on Instagram subject to an age restriction (18+) illustrates the difficulty of balancing between protecting freedom of expression and safeguarding the dignity, privacy, and reputation of individuals. In our assessment, the approach taken is inadequate given the scale and specific harms associated with NSII.

Generative AI tools have dramatically lowered the barrier to producing realistic sexually explicit material depicting individuals without their knowledge or consent. While the creation of deepfakes once required advanced technical expertise, widely available applications now enable face-swapping or the generation of entirely synthetic footage.

Our findings indicate that the overwhelming majority of victims are women and girls. Such material is frequently weaponized for harassment, blackmail, non-consensual pornography distribution (commonly referred to as 'revenge porn'), or the professional discrediting of the victim. The resulting harms are not merely reputational, but also psychological, professional, and social.

The dimension of social harm should be central to any assessment of this type of material. Even where content does not depict an explicit sexual act, its realistic character and sexualizing context can produce concrete, measurable harm in both the digital and offline spheres. An approach based solely on the analysis of visual content (asking only whether a sexual act is visible) is therefore insufficient.

The core issue with NSII is not nudity per se, but the absence of consent to the use of a person's image in a broadly sexual context. We accordingly recommend that platform policy explicitly classify NSII as a distinct category of violation, irrespective of whether the content depicts a full sexual act. Putting consent in the center is necessary for recognition of individuals' right to exclusive control over their own image.

In a case such as this, the operative question is not 'Is a sexual act visible?' but rather 'Did this person consent to being depicted in a sexualizing manner?' and, more broadly, 'Did this person consent to being depicted at all?' The case description indicates that no such consent existed, although there is no indication that the depicted individual personally sought removal of the content.

One of the most significant practical challenges is establishing whether a person has in fact consented to the use of their image. Platforms frequently lack direct contact with the individual depicted. We therefore recommend the adoption of a presumption of harm principle, which in practice means that:

1. A credible statement by the individual that they did not consent should be sufficient to trigger protective action.
2. The burden of proof should not fall on the victim.
3. Where doubt exists, decisions should be made in favor of protecting the privacy and image rights of the person making the report.

We further recommend the introduction of a simplified reporting procedure for private individuals, enabling swift identity verification and submission of a non-consent declaration, resulting in the temporary blocking of content pending full review. This requires the implementation of technically secure mechanisms for attaching proof of identity, as well as clear processes for informing complainants of the status of their case.

Enhanced protection for private individuals and response speed

Case 2026-021-IG-RA concerns a private individual, which warrants particular consideration. Private individuals should benefit from a heightened standard of protection regarding their image and privacy. It must be emphasized that non-consensual sexualizing content should not enjoy the protections afforded to parody, satire, or artistic expression. The starting point for any assessment of legality and harm should be whether the individual depicted has given or withheld consent.

In NSII cases, a rapid response is essential to limiting harm. Every delay increases the number of views, copies, and secondary publications. The failure to prioritize reports concerning potential non-consensual sexualization represents a systemic problem, as illustrated by this case, where multiple reports and appeals did not result in human review.

Content reported as NSII, or flagged by automated systems as potentially harmful intimate imagery with a high likelihood of virality, should be automatically routed to an expedited verification pathway staffed by trained moderators, with particular attention to the protection of private individuals. Pending human review, such content should be temporarily restricted rather than left publicly accessible.

Age-gating insufficiency and freedom of expression

In the discussed case, Meta determined that restricting visibility to users aged 18 and over constituted an adequate response. This approach is based on a flawed assumption, as The harm caused by NSII does not arise from children accessing the content. The very existence and distribution of the material is harmful for the targeted individual.
Age-gating:

1. Does not protect the victim's reputation or psychological well-being.
2. Does not prevent further copying and redistribution of the content.
3. Does not eliminate the risk of harassment and blackmail.

Age-gating may be an appropriate tool for consensually produced erotic content, but it should not be the default response in NSII cases. Where non-consensual intimate imagery is involved, the appropriate response is removal, not restriction.

NSII constitutes a specific type of violation that need not be classified under a narrow understanding of 'pornography.' The subjective experience of harm by those affected is of paramount importance. We therefore recommend a shift toward the category of AI-generated intimate imagery without consent as the operative framework for platform policy.

Meta has referenced its 'fundamental commitment to expression.' Freedom of expression is indeed a foundational value; however, in line with international human rights standards, it is not absolute. Content that involves impersonating or sexualizing a private individual without consent makes no meaningful contribution to public discourse, is abusive in character, and generates real harm to individuals and to society. Restricting such content through moderation and, where appropriate, removal is justified and proportionate as a measure to protect the rights of others.

Final recommendations

On the basis of the above analysis, we propose that the Oversight Board considers issuing recommendations that would require Meta's digital platforms to:

1. Establish a distinct violation category for non-consensual sexualizing imagery (NSII/NCII), separate from existing adult content standards.
2. Adopt removal (not age-gating) as the default response to NSII, in line with the principle that harm arises from the content's existence.
3. Create a priority moderation pathway for NSII reports, with expedited human review by trained moderators with expertise in synthetic media and digital gender-based violence.
4. Ensure that content flagged by automated systems as potentially harmful intimate imagery with a high likelihood of virality, or reported as non-consensual pornography, is temporarily restricted and immediately escalated to human review.
5. Maintain a number of human moderators proportionate to the volume of users across relevant languages, to guarantee timely decision-making.
6. Provide moderator training in the recognition of AI-generated and synthetic media, and in the context of digital violence and NSII victimization.
7. Implement technical tools to prevent re-upload of removed NSII content, including perceptual hashing.
8. Publish transparent statistics on NSII-related reports, enforcement actions, and outcomes.
9. Develop partnerships with civil society organizations specializing in responses to digital violence and online abuse.
10. Introduce simplified, secure reporting mechanisms for private individuals, including identity verification and non-consent declaration processes, with clear status updates for complainants.

We hope this comment will contribute to making Meta's platforms safer for all users and remain available to answer any additional questions the Board may have.

CEE Digital Democracy Watch
Mokotowska 43/104, 00-551 Warsaw, Poland
www.ceeddw.org
KRS 0001090110 | EU Transparency Register 114284692189-05