Non-consensual AI sexualised impersonation, 2026-021-IG-RA

Submission by Stacey Kelly-Maher

Prevalence and harms of image-based abuse:

Image-based is a deeply gendered and increasingly prevalent issue. My Image My Choice found there were over 270 thousand videos on the top 40 sites dedicated to AI-generated image-based abuse, a 3,000% increase from 2019. Sippy et al. (2024) surveyed 1,403 UK adults to understand exposure to and perceptions of AI-generated imagery and found that 18.8% of participants had been exposed to AI-generated intimate image abuse. In the 2023 State of Deepfakes report, 98% of AI-generated videos online were found to be sexualised and 99% of those depicted in these sexualised AI-generated videos were women.

Experiencing image-based abuse can have significant impacts on survivors, including their physical safety, emotional wellbeing, economic status, and more. This impacts marginalised communities in distinct and compounding ways. For example, Glitch has found that Black women are disproportionately likely to experience AI-generated image-based abuse, and WIRED estimated that 5% of imagery generated by Grok featured women who were stripped from or made to wear religious or cultural clothing.

The Oversight Board has in a previous case stated that deepfake intimate images have a disproportionate impact on women and girls, their rights to privacy, and protection from mental and physical harm. The findings of 2024-007-IG-UA and 2024-008-FB-UA found that "Given the severity of harms, removing the content is the only effective way to protect the people impacted."

Review process and policy enforcement:

In case 2026-021-IG-RA, multiple days passed before human review, and a timely response is vital when it comes to image-based abuse. Potential image-based abuse must be prioritised for human review, and it should not require contact with the Board for content to be reviewed. Decreased turnaround of review/takedowns must be prioritised in cases of image-based abuse, and this will be in line with the UK Government's proposed 48 hour maximum for the takedown of image-based abuse. This would be an absolute maximum, and Meta should aim for shorter timeframes than this as every moment such imagery is available on platform, it is available to be seen and saved by

others. At a minimum, while this content is awaiting human review, it should be downranked to limit its further dissemination.

Once reviewed and identified as AI-generated image abuse, this content should be added to a hashing database such as StopNCII to prevent the upload or re-upload of this content across platforms. Sippy et al.'s (2024) research indicates that 87.3% of UK adults support the ban or suspension of users who distribute AI-generated imagery that may be harmful, and 82.4% believed social media platforms should make it easier to report harmful AI-generated imagery and request content removal.

The Bullying and Harassment Community Standard states that as a Tier 1 universal protection for everyone, everyone is protected from "derogatory sexualised photoshop or drawings". This means this post was by its nature due for removal, and I would also recommend that the language in this Community Standard is updated to "non-consensual synthetic intimate imagery" or similar. This would expand beyond "photoshop or drawings" to clarify that any synthetic content would be unacceptable, and it would go beyond the indeterminate "derogatory" to be clear that regardless of intention, no non-consensual image-based abuse can be allowed on the platform. This is in line with previous Board findings. I would also recommend that Meta's position on AI-generated image-based abuse is clarified in the NCII section of the Bullying and Harassment topic as this is not mentioned currently.

The Board decision relating to 2024-007-IG-UA and 2024-008-FB-UA found that labelling is not appropriate as the harm comes from sharing and viewing, and this equally applies in this case where age-gating may limit the spread, but ultimately does not address the core harms and leaves the imagery available to a significant audience.

Further recommendations:

In Ofcom's Deepfake Defences paper, they share potential enforcement actions such as issuing warnings/strikes to users who have broken such guidelines, taking down content, suspending or removing users, and labelling content where there isn't a clear breach. Their guidance on improving women and girls' online safety sets out good practice steps including: consulting with subject-matter experts around setting policies such as this, training employees responsible for setting policies on online gender-based harms, requiring evidence of consent from those depicted in intimate content prior to upload, and adding deterrence messaging to the upload process. I would also recommend that if these images are believed to be consensual, this consent would still have to be understood to be ongoing and if at any stage someone wishes to be able to revoke consent in imagery that depicts them, they should be able to do so.

Ofcom note that "it is important for providers to recognise that user reporting relies on survivors and victims of online gender-based harms, and that reporting processes are time-intensive and risk re-traumatising survivors and victims". They state that users can be encouraged and enabled to report by making these reporting systems accessible, transparent, easy-to-use, and accounting for the dynamics of online gender-based harms. Foundational steps in this guidance and the related Codes include acknowledging receipt of complaints, providing indicative timeframes, and setting out information about how the complaint will be handled. Good practice steps to go further could be quick exit buttons, allowing report tracking, allowing users to give feedback on reporting processes, signposting to relevant supportive materials and organisations during reporting, and establishing trusted flagger programmes.

Finally, Meta is a signatory of the IBSA Principles for combatting image-based sexual abuse which highlight trauma-informed approaches and investing in resources and tooling to ensure the rapid processing of reports. For further resources on improving reporting processes, Chayn and End Cyber Abuse have created the Orbits guide which includes detailed principles for trauma-informed approaches to technology-facilitated abuse and an audit template for reviewing platforms according to these principles.