Where Governance Fails on AI-Generated Nonconsensual Intimate Imagery

Sara Alterazi & Patrick Gilmartin – 5/11/26

A mom posts a cute video of her baby online. She and her husband have been struggling financially, so she thought TikTok may be the answer. She often sees people rise to fame overnight and become wealthy. At first, it seemed harmless: family members see the video and comment, friends like it, strangers may even save it. But the same variable that makes social media fun also leads to risk. With the advancement of AI, simple videos posted on tiktok can be turned into sexual images with a few taps on a phone. This is why AI -generated nonconsensual imagery is not just a deepfake problem; it is a governance problem.

The TAKE IT DOWN Act uses the term digital forgery to describe intimate images that were made through software, machine learning, or artificial intelligence that look like a real person (TAKE IT DOWN Act, Pub. L. No. 119-12, § 2(a), 2025). Consent is also defined as a clear, voluntary agreement made without force, fraud, pressure, or manipulation (TAKE IT DOWN Act, Pub. L. No. 119-12, § 2(a), 2025). These definitions matter because AI- generated nonconsensual intimate imagery is not just fake content, but content that uses a real person's identity without permission, which results in reputational damage and the false appearance of authenticity. This issue is even more serious when minors are involved.

In December 2025, a Tennessee high school student learned, through an anonymous Instagram message, that sexualized AI-generated images of her were being shared on Discord. These images were generated using ordinary, nonsexual photos, including her yearbook picture and images pulled from social media. They were altered, using AI tools connected to xAI's Grok and third-party "undressing" apps. Police later found that the same offender had created similar images of other girls, and in March, 2026, a three-person joint lawsuit was filed against the company, arguing that its tools helped make such abuse possible (Huo).

Researchers Michelle L. Ding, Harini Suresh, and Suresh Venkatasubramanian describe scenarios like this as a "whack-a-mole" problem. When policymakers or platforms focus only on removing one image, banning, one app, or punishing one user, the larger system remains untouched. Their framework shows that AI-generated nonconsensual imagery is supported by an ecosystem of tools that help create, distribute, discover, and monetize the abuse. This means governance cannot only be reactive. If the response begins after an image is already circulating, the law is constantly chasing harm that the technology has already made easy to produce (Ding et al.). More offensive measures are a must.

In 2025, the National Center for Missing & Exploited Children received more than 1.5 million CyberTipline reports indicating connections between generative AI and child sexual exploitation ("The Work Never Stops", Vaughan). These reports included AI-generated images and videos, as well as chats, files, and other activity where generative AI played some role. This illustrates the pervasiveness of the issue and sheds light on the fact that this problem is no longer

limited to isolated deepfake scandals or celebrity targets. It is becoming part of the ordinary architecture of online exploitation. A child's school photo, a teenager's Instagram post, or a family video can become source material for abuse, and the victim may only find out after the image has already moved through private servers and been shared in encrypted group chats. For AI governance, this means the central question cannot only be how to remove harmful content once it's reported. The harder question is how to design accountability across the systems that make the abuse easy to create, easy to circulate, and difficult to trace.

The TAKE IT DOWN Act responds to one part of that problem by requiring covered platforms to remove nonconsensual intimate images after receiving a valid request (TAKE IT DOWN Act, Pub. L. No. 119-12, § 2(a), 2025). That is important, especially for victims who otherwise have to beg each website to act. But removal alone cannot carry the entire policy response. Platforms should also be required to search for and remove identical copies, prioritize cases involving minors, preserve evidence for law enforcement, and make reporting tools clear enough for ease of use. A governance system that waits until the image is already everywhere is not preventing harm; it is cleaning up after a system that was allowed to function in such a manner in the first place. Instead, we must exterminate the issue at its root.

Companies that build image-generation systems should not be able to treat misuse as a separate platform problem after their tools have already enabled the harm. If a model can generate sexualized images of identifiable people from ordinary photos, the company should have a duty to test for that risk before release, block prompts and uploads designed to sexualize real people, and limit integrations with third-party apps that market themselves distastefully. This is especially important when the system allows users to upload a real person's face or body as the input. In those cases, the company is not just providing a neutral creative tool; it is creating the conditions for identity-based sexual exploitation. AI governance should therefore require development-stage safeguards, abuse testing, audits records, and clear accountability when companies ignore foreseeable misuse.

This can be done through layered safeguards, not just a warning screen or a terms-of-service rules. Before release, companies should red-team their models against predictable abuse cases. At the product level, systems should use input and output classifiers that detect when a user is trying to combine an identifiable person with nudity or sexual prompts, then block the request before an image is generated. Companies should also limit API access for high-risk image tools, log abuse patterns, rate-limit suspicious behavior, and cut off third-party apps that repeatedly route users toward sexualized functions. After content is generated, tools like watermarking, provenance metadata, and perceptual hashing can help platforms trace where synthetic images came from and find reuploads more quickly.

Unfortunately, no single safeguard will put a stop to the abuse. Bad actors will find a way to commit these criminal acts, regardless of what's put in place. It is the duty of AI companies, therefore, to make generation of such content as difficult as possible. To ensure damage control and mitigation are at peak capability. This can be achieved with layered controls that make exploitation harder to produce, easier to detect, and less profitable to distribute.

Works Cited

Ding, Michelle L., et al. "How to Stop Playing Whack-a-Mole: Mapping the Ecosystem of Technologies Facilitating AI-Generated Non-Consensual Intimate Images." *arXiv*, 4 Feb. 2026, arXiv:2602.04759.

Jingnan, Huo. "Tennessee Teens Sue Elon Musk's Xai over AI-Generated Child Sexual Abuse Material." *NPR*, NPR, 17 Mar. 2026, www.npr.org/2026/03/16/nx-s1-5749490/xai-elon-musk-sexualized-images.

"The Work Never Stops: A First Look at NCMEC's 2025 Data." *National Center for Missing & Exploited Children*, 31 Mar. 2026, www.missingkids.org/blog/2026/the-work-never-stops-first-look-at-ncmecs-2025-data.

Tools to Address Known Exploitation by Immobilizing Technological Deepfakes on Websites and Networks Act ("TAKE IT DOWN Act"), Pub. L. No. 119-12, § 2(a), 139 Stat. 55 (2025).