# AI-generated Image Abuse: Closing the Accountability Gap

*Policy Brief*

AI FORENSICS

## The Evidence

Between **December 25, 2025 and January 1, 2026**, AI Forensics collected 50,000 tweets prompting Grok for image generation, and 20,000 AI-generated images from @Grok on X with no filter for specific content. Within this sample:

- 53% of images showed individuals in minimal attire—81% were women
- 2% depicted persons appearing under 18[1], including children under 5[2]
- 6% depicted public figures (350+ individuals, one-third politicians)

95% of @Grok mentions were replies to others' posts. Users weaponized the tool against women by requesting sexualized manipulations of their photos without victims' consent. 83% of requesters appeared male. It is important to note that users did not need to circumvent the system, Grok generated the images with simple and straightforward user prompts.

After January 9, X restricted image generation to paid subscribers. Before this change, verified users represented 22% of generations; after, they represented virtually all. X monetized the restriction rather than removing the capability.

**Inconsistent Patches**

Between December 25, 2025 and January 1, 2026, around half of the images generated by @Grok in our sample depicted people in minimal attire. By January 13-14, 2026, this fell below 10%, suggesting some safeguards were implemented on X.

[1] Not necessarily in minimal attire.
[2] Details on our methodology can be found in our data report:https://www.aiforensics.org/work/grok-unleashed

However, as reported in Wired[3], *Grok Imagine* operated by xAI, was used to produce extremely graphic images and videos that are vastly more explicit than images created by @Grok on X such as content involving full nudity, and/or sexual acts.

**As of January 20th, Grok can still be used to generate sexualized images of individuals.**

We analyzed two thousand user conversations with Grok. The overwhelming majority of analyzed content depicted nudity and/or individuals in sexual activity alongside deepfakes of political figures and synthetic media depicting individuals suspected of being minors. These have been flagged to French regulators and *Point de Contact*[4], a French Trusted Flagger under the Digital Services Act (DSA) for Child Sexual Abuse Material (CSAM) identification and reporting.

## The Case: Grok on X

### Platform - AI Split

X is a social media platform; Grok is an AI chatbot developed by xAI. Though different products, they belong to the same entity[5]. X chose to embed Grok on its platform, where AI-generated outputs become public by default, unlike standalone chatbots where users may actively choose to share outputs.

The convergence of platforms and generative AI systems presents a unique regulatory challenge. EU law treats platforms and AI systems differently, leading to inconsistencies in how content liability and moderation responsibilities are understood and assigned.

- **The Digital Services Act (DSA)** governs platforms hosting user content.
- **The AI Act (AIA)** treats AI systems as products requiring pre-deployment safety measures focusing on training data and model design, not real-time outputs. Generative AI outputs involve both user prompt and model parameters set by the provider. For instance, Grok does not merely host content, but also generates it.

### DSA Obligations

Under Articles 34-35 of the DSA, as a Very Large Online Platform (VLOP), X, is required to assess and mitigate systemic risks, including the dissemination of illegal content, negative effects on fundamental rights and the rights of the child, gender-based violence, and manipulation affecting public security and minors. This includes assessing how the platform's subsequent features, such as the ability to generate media via Grok, aggravate the severity and probability of these risk categories.

---

[3] https://www.wired.com/story/grok-is-generating-sexual-content-far-more-graphic-than-whats-on-x
[4] https://www.pointdecontact.net/
[5] https://www.bbc.com/news/articles/ceqjq11202ro

In its risk assessment reports, X acknowledges generative AI as a significant source of inherent risk regarding illegal content, disinformation, and fundamental rights without citing Grok explicitly[6]. Its mitigation measures include: proactive detection of violative content, media hashing to detect known CSAM (integrated with StopNCII.org to use digital fingerprints ((hashes)) to proactively block the spread of non-consensual intimate images), and labels for synthetic and manipulated media (SAMM) that may confuse and deceive people and lead to harm.

Given that Grok is embedded into X, The European Commission's data retention order for Grok can help investigate whether X's systemic risk assessment and mitigation measures have proved sufficient. However, Grok's web and mobile apps currently are not covered by the DSA. This enables regulatory arbitrage. As DSA enforcement pressure increases on X, harmful capabilities remain accessible on Grok.com, as we reported.

### AI Act Obligations

Under the AIA, GPAI providers have transparency obligations. The AIA does not govern the outputs of GPAI models, it requires the synthetic content to be watermarked and machine-detectable. However, if Grok qualifies as a GPAI model with systemic risk—arguably, its reach via X can be considered a decisive factor beyond computational thresholds—additional assessment and mitigation duties for systemic risks apply (relatively similar framing to systemic risks in the DSA).

xAI signed the GPAI Code of Practice safety and security chapter, therefore committing to identify "risks from illegal, violent, hateful, radicalising, or false content, including risks from child sexual abuse material (CSAM) and non-consensual intimate images (NCII)" outlined in Appendix 1.4. The Code of Practice is a voluntary initiative and not legally binding. However, the AI Office can open an investigation on the basis of these commitments. This would help prevent the regulatory arbitrage and accountability gap stemming from GPAI deployment architecture.

### Liability and Moderation

Under the DSA, platforms are not held liable for user content. The framework assumes harmful content comes from users, with platforms as intermediaries required to promptly remove illegal content once notified. Platforms also have the right to moderate content that does not respect their terms and community guidelines (even if it's not forbidden under the applicable law).

Platform moderation and AI chatbot moderation operate according to significantly different logics. Platform moderation is reactive: It addresses user-generated content that already exists and can be removed once flagged. AI chatbot moderation must be both anticipatory and reactive. This requires a dual approach: "moderating behavior" (pre-deployment measures like

[6] https://transparency.x.com/content/dam/transparency-twitter/dsa/2025-x-dsa-sra-summary-report.pdf

training data curation and fine-tuning) and "moderating content" (real-time filtering through classifiers and blockers)[7]. AI systems are simultaneously generating and moderating their outputs- a paradigm shift from platform intermediary liability that current regulations have not fully addressed.

X states that AI-generated content is subject to X's rules and those are enforced irrespective of how content is produced[8]. Similarly, xAI's terms state users "own and are responsible for" AI outputs while simultaneously claiming broad rights to all user data[9]. This externalizes liability while retaining commercial benefit. The AIA places obligations on GPAI providers but does not establish civil liability for harmful outputs. The revised Product Liability Directive may apply but remains untested. The proposed AI Liability Directive is stalled. Victims currently have no clear path to remedy.

The question of liability for generative AI has broad implications, including around copyright. While it is not possible to eliminate all harmful outputs from generative AI systems due to their non-deterministic nature, the concept of moderation can extend to chatbots[10]. Reactive measures such as system prompts, filters, blockers, and classifiers can be evaluated—both the decision to implement them and their effectiveness. This approach could help foster accountability for similar cases without settling these questions entirely.

The following factors provide strong grounds to impose such obligations, particularly in this context:

1. **No circumvention required**: The system worked as intended. The provider cannot claim lack of foreseeability. This constitutes foreseeable misuse.
2. **Internet connectivity**: Real-time access to current events and public figures increases the potential for targeted harassment.
3. **Platform embedding**: X integration creates distribution infrastructure. Generated content is one click from 500+ million users.
4. **Unified ownership**: xAI and X share ownership. Coordination between GPAI provider and platform is not merely possible but occurring. This is not a third-party integration.

### Beyond "sexually-explicit" definitions

Consensual AI-generated images are created with clear permission from the person depicted. Non-consensual images involve generating or altering someone's likeness without their approval—frequently in ways that are sexualized, humiliating, or damaging. Without consent, this technology becomes a vehicle for violating autonomy and dignity. The harm is real regardless of whether the image is "fake."

---

[7] https://www.aiforensics.org/work/governing-ai-search
[8] https://transparency.x.com/content/dam/transparency-twitter/dsa/2025-x-dsa-sra-summary-report.pdf
[9] https://x.ai/legal/terms-of-service
[10] https://aiforensics.org/uploads/Governing_AI_Search.pdf

Image-based sexual abuse predominantly impacts women, who account for 90% of victims. According to UNESCO, 58% of women worldwide have experienced technology-facilitated gender-based violence. Deepfake content has increased 550% since 2019—99% of it targeting women and girls[11]. Legal frameworks often focus narrowly on "sexually explicit" content. But harm occurs across a broader spectrum of intimate images—whether women in bikinis or AI-generated images removing a Muslim woman's hijab. Current definitions fail to capture how violence and control operate through image manipulation, particularly for women facing intersecting forms of discrimination.

### Legislative Gaps

The proposed Directive on Combating Violence Against Women and Domestic Violence asks member states to ensure that Article 7(b): "producing or manipulating and subsequently making accessible to a multitude of end-users, by means of information and communication technologies, images, videos or other material, making it appear as though another person is engaged in sexual activities, without that person's consent" is punishable by criminal law.

The requirement of "subsequently making accessible to a multitude of end-users" should not be a precondition for criminalizing non-consensual image-based abuse. AI chatbots make available "share" features that render conversations indexed and accessible even when not shared on a platform.

Additionally, GPAI model providers and deployers can monitor the use of their tools and detect this kind of activity. Their monitoring and reporting obligations should be mandatory regardless of whether they are designated as a GPAI model with systemic risk under Article 54 of the AI Act.

## Recommendations

**For the European Commission**

1. Investigate whether X's systemic risk assessment adequately addressed the integration of Grok and whether mitigation measures proved sufficient.
2. Clarify that GPAI providers embedding their models in VLOPs bear joint responsibility for systemic risk mitigation.
3. Issue guidance on how the DSA and AI Act interact when AI systems are deployed on regulated platforms.

**For the AI Office**

1. Open an investigation into xAI's compliance with GPAI Code of Practice commitments, particularly regarding CSAM and NCII risk identification.

[11] https://donestech.net/en/noticia/2023-state-deepfakes-home-security-heroes

2. Assess whether Grok meets the threshold for GPAI with systemic risk given its distribution via X.
3. Develop standardized requirements for reactive moderation measures in GPAI systems with public-facing deployment.

**For Member States**

1. Ensure transposition of the Violence Against Women Directive does not condition criminalization on public distribution.
2. Extend mandatory monitoring and reporting obligations to all GPAI providers, not only those designated as systemic risk models.

**For the Co-legislators**

1. Ensure product liability frameworks account for the dual nature of AI systems as both content generators and content moderators.

## Appendix: Timeline

In August 2025, xAI introduced the so-called "Spicy Mode" that allows users to generate adult content (including full nudity) on Grok.com. The conversational robot @grok on X, capable of generating images, was increasingly asked by users to undress individuals (e.g. "@Grok, put her in a bikini"). This escalated throughout December 2025.

**Dec 28**. Grok posted[12] an "apology" acknowledging it generated sexualized images of two young girls (ages 12-16) — a failure in safeguards that violated ethical standards.
**Jan 2nd**. French authorities opened an investigation[13] on the proliferation of sexually explicit deepfakes generated by artificial intelligence platform Grok on X following French lawmakers Arthur Delaporte and Eric Bothorel.
**Jan 3.** Elon Musk issued statement[14] saying "*Anyone using Grok to make illegal content will suffer the same consequences as if they upload illegal content.*"
**Jan 5**. AI Forensics, a European NGO, published[15] the first large-scale analysis of non-consensual sexualized images generated by Grok on X.
**Jan 7**. Wired and AI Forensics reported[16] on the capability and usage of grok.com to generate fully explicit pornographic content.
**Jan 8.** The European Commission extended a data retention order for X to preserve all internal documents and data related to Grok until the end of 2026.
**Jan 9.** X restricted image generation/editing to paid subscribers only.
**Jan 12.** Ofcom launched an investigation into X over Grok sexualised imagery.

[12] https://www.techpolicy.press/the-policy-implications-of-groks-mass-digital-undressing-spree
[13] https://www.politico.eu/article/france-lawmaker-investigate-deepfakes-women-stripped-naked-grok-x
[14] https://x.com/elonmusk/status/2007475612949102943
[15] https://www.aiforensics.org/work/grok-unleashed
[16] https://www.wired.com/story/grok-is-generating-sexual-content-far-more-graphic-than-whats-on-x

**Jan 12.** Malaysia and Indonesia blocked Grok.

**Jan 14.** California State Attorney General launched an investigation into xAI, Grok over undressed, sexual AI Images of women and children.

**Jan 15**. X declared[17] having implemented measures to prevent @Grok from being used to create intimate images of people.

[17] https://x.com/Safety/status/2011573102485127562