Belen Bringas
15/06/2026
AIPL Externship

Policy Brief: Regulating Harm: AI-Generated Non-Consensual Deepfakes as Gender-Based Violence in the EU and UK

1. Executive Summary:

AI-generated non-consensual deepfakes are a growing form of image-based abuse and gender-based violence. Generative artificial intelligence (AI) allows for the creation of intimate images without consent, and it affects women and girls most, causing psychological, reputational, and professional harm. Despite increasing public attention and recent governance developments, gaps remain. This brief addresses the following question: How do the European Union's AI Act and the United Kingdom's Online Safety Act regulate AI-generated non-consensual deepfakes as a form of gender-based violence? Through qualitative thematic textual analysis, five analytical themes are identified, and the findings reveal that the AI Act approaches harm from a risk-based perspective, while the Online Safety Act uses an offense-based approach to define harm. Both Acts also include preventive and reactive measures but fail to explicitly recognize AI-generated non-consensual intimate images as gender-based violence. Drawing on feminist governance theory, this brief provides recommendations for the recognition of non-consensual deepfakes as gender-based violence, expanded responsibilities in the AI life cycle and platforms, an emphasis on victim-centered measures, a preventive governance approach, and robust international cooperation to combat the transnational nature of online harms.

2. Background:

Late December of 2025 marked an unprecedented moment when social media platform X, formerly known as Twitter, was flooded by more than a million AI-generated images of sexualized women produced by the platform’s AI chatbot, Grok. Requests asking it to remove clothing and alter images into sexualized positions resulted in more than 4.4 million images being produced, with at least 41% featuring women in just nine days (Conger et al., 2026). While nudifying tools exist and remain problematic, the competition is slim when considering the distribution of the content that comes with an AI bot integrated into a popular social media platform, a sentiment shared by one of the godfathers of AI, Yoshua Bengio, who claimed that artificial intelligence and its industry had become too unconstrained and lacked the necessary

technological and societal constraints. Following the scandal, X announced restrictions on Grok that would prevent it from manipulating images of real people in revealing clothing (Milmo, 2026). The international response has been overwhelming, with the European Union reaching an agreement in May of 2026 to amend the AI Act by prohibiting AI systems that produce non-consensual intimate images, such as nudification apps (European Commission, 2026). This recent development reflects the growing recognition amongst policymakers that current governance frameworks must better address AI-enabled sexual abuse.

This policy brief focuses on how the EU AI Act (AIA) and the UK Online Safety Act (OSA) regulate AI-generated non-consensual deepfakes as a form of gender-based violence. Deepfakes are understood as "an image or recording that has been convincingly altered and manipulated using an AI technology to misrepresent someone as doing or saying something that they did not do or say" (Okolie, 2023, p. 5). These AI-generated non-consensual deepfakes fall under the category of image-based sexual abuse (IBSA), which is a form of gender-based violence (GBV). Image-based abuse is defined as "non-consensual creating, taking or sharing of intimate images or videos, including altered or manipulated media" (Rigotti et al., 2024, p. 1473). AI-generated non-consensual deepfakes create disproportionate harm to women; for example, in one dataset of 85,000 deepfake videos, 90 percent portrayed women in non-consensual sexual content (Akter & Ahmed, 2025, p. 7). Despite their rapid spread, governance frameworks have struggled to keep pace, leaving victims without the necessary protective measures (Rigotti et al., 2024, p. 1473).

3. Problem Statement:

The debate around AI-generated non-consensual deepfakes as image-based sexual abuse and the exact definitions are rather wide, but the consensus is that they are a form of gendered IBSA that is used as a tool to humiliate, silence, subjugate, and control women. It is important to mention that nudification apps that are powered by AI to undress women and perform face swaps are also included in IBSA, as they allow the rapid creation of non-consensual intimate images (NCII) (Karagianni and Doh, 2024; Laffier and Rehman, 2023; Okolie, 2023; Williams, 2025). There is a combination of factors that allows AI-generated non-consensual deepfakes to be considered as GBV, which includes their disproportionate harm towards women and the kind of harm that is produced and amplified by generative AI. The malicious use of this technology to produce NCII

can be used for harassment, identity manipulation, blackmail, and privacy violations, which can result in a variety of harms (Casagrande and Spaticchia, 2025; Prado, 2025). Some of these include, but are not limited to, psychological distress such as anxiety and depression, which can lead to suicide cases, reputational and professional damage, self-censorship, and withdrawal from public as well as digital spaces. These harms represent a clear violation of identity, privacy, dignity, and personality rights. Feminist scholarship is especially strong in this regard, as AI-generated non-consensual deepfakes disproportionately target women and normalize women's sexualization. This form of IBSA reproduces and strengthens patriarchal structures and norms, reinforces misogynistic power structures, and upholds systematic gender inequality.

4. Key Findings:

This brief used qualitative theory-driven thematic textual analysis to examine how the EU AI Act and the UK Online Safety Act regulate AI-generated non-consensual deepfakes as a form of gender-based violence. The primary material consisted of two legal texts: the European Union's AI Act and the United Kingdom's Online Safety Act. Initial codes included gendered harm, image-based sexual abuse, victim protection, prevention, accountability, platform responsibility, developer responsibility, transparency, risk classification, coercion, enforcement, intersectionality, and omissions. These codes were subsequently grouped into five analytical themes: how harm is defined, allocation of responsibility, preventive versus reactive governance, recognition of gendered harm, and regulatory omissions.

4.1 How Harm is Defined:

The EU AI Act mainly defines harm through the concepts of risk, fundamental rights, health, and safety. Therefore, the framework conceptualizes harm as a regulatory risk that needs to be managed. AI-generated non-consensual deepfakes are not identified explicitly as a form of harm but are indirectly captured under violation of fundamental rights, a broader category of harm. The UK Online Safety Act, on the other hand, defines harm through diverse concepts such as harmful communications, illegal content, harassment, threats, intimate image abuse, and psychological harm. Thus, harm is understood through offense-based terms, which are more concrete than the AIA's more abstract risk-based approach. An important difference is that the OSA makes a direct connection between harmful online content, identifiable victims, and criminal conduct.

4.2 Allocation of Responsibility:

In the AIA, responsibility is distributed between several actors throughout the whole AI life cycle, such as deployers, providers, importers, distributors, operators, and authorized representatives. Hence, responsibility is allocated to all the actors involved in the development, deployment, and operation of AI systems. In comparison, the OSA allocates responsibility to platforms, users, regulators, service providers, and individuals who disseminate harmful content. The key distinction between the two frameworks is that the OSA spreads the responsibility through multiple actors in the online ecosystem, while the AIA focuses on the technology developers.

4.3 Preventive versus Reactive Governance:

The AIA heavily relies on preventative governance as it emphasizes identification, mitigation, and anticipation of risks before they occur. Harm prevention is also embedded in the development, deployment, and monitoring of AI systems. The OSA uses both preventative and reactive approaches, as the framework has the goal of reducing risk before it occurs, but also provides enforcement mechanisms for after harmful content is created or shared.

4.4 Recognition of Gendered Harm:

There is a limited recognition of gendered harms in the AIA as it does not explicitly reference women, GBV, IBSA, or AI-generated non-consensual deepfakes. Gendered harms are consequently indirectly addressed through broader categories of harm such as bias and discrimination. The OSA, on the other hand, acknowledges several forms of abuse that come from AI-generated NCII, but despite this recognition, the framework does not characterize these harms as GBV or recognize the disproportionate harm for women and girls.

4.5 Regulatory Omissions:

While the AIA defines deepfakes under transparency obligations, it neither classifies AI-generated non-consensual deepfakes as prohibited systems nor as high-risk systems. Other omissions include the lack of references to GBV, IBSA, platform accountability, and victim-centred measures. It should be noted that the May 2026 agreement to amend the AI Act would prohibit AI-generated NCII, including nudification applications. However, this amendment had not yet entered into force at the time of analysis, nor had it been formally adopted or been in force as of June 2026, when this brief was written. The OSA, in contrast, does

criminalize AI-generated non-consensual deepfakes and has clear rules for platform moderation and content distribution, but it does not frame deepfakes as GBV. There is also limited attention placed on the development of AI systems capable of producing NCII. Additionally, there are omissions regarding cross-border enforcement and global circulation once the content is distributed online.

5. Policy Recommendations:

The recommendations are guided by feminist governance theory, which emphasizes that governance is not neutral and that institutions, laws, and policy are shaped by power relations that can reinforce structural inequalities (Sawer et al., 2023, pp. 1–2). This framework is particularly relevant for the analysis of AI-generated non-consensual deepfakes as a form of gender-based violence because feminist scholarship has argued that these forms of image-based sexual abuse disproportionately affect women. Feminist scholars have claimed that both systems and governance models often universalize users and ignore the disproportionate harm that marginalized groups experience by overlooking how aspects such as gender, race, sexuality, and other identities shape the experience of online harm (Shade and Shepherd, 2025, pp. 7–8).

5.1 EU Recommendations:

The findings reveal that the original AIA regulated deepfakes through transparency provisions and, despite the recent amendment, show a significant improvement by prohibiting AI systems capable of producing NCII, such as nudification apps; however, several gaps remain. There are other forms of generating non-consensual deepfakes, such as face-swapping, that should also be addressed in this amendment. Furthermore, feminist governance emphasizes that regulation should address the structural conditions that enable abuse rather than focusing on a specific category of technology. More work is needed to understand the underlying conditions that lead to AI-enabled sexual abuse, and future reforms should cover a broader range of techniques used to create NCII. The findings also show that the AIA lacks explicit victim protection measures and that, despite the risk-based approach and compliance obligations, there is limited focus on the impact of AI-generated non-consensual deepfakes on victims' lives. A feminist governance lens shifts the focus to the experiences and needs of victims in order to protect them through governance structures. Future reforms are therefore necessary and should include mechanisms for content removal, reporting mechanisms, and support services for those affected.

5.2 UK Recommendations:

The OSA places significant responsibility on platforms through complaint procedures, risk assessments, and moderation obligations, but places less focus on the developers and providers of AI systems capable of producing AI-generated NCII. Following a feminist governance perspective, all actors involved in the AI life cycle, as well as digital platforms, should share responsibility. The recommendation is therefore to take developers, providers, and deployers into consideration and create obligations for them to prevent the production of non-consensual deepfakes. The findings on the OSA demonstrate a combination of both preventive and reactive governance, but with a strong focus on criminal offenses and post-harm enforcement. Preventing harm before it occurs is key to feminist governance, and while criminalization is important, preventive obligations should be equally important. Future amendments should consider risk mitigation and safety measures before harmful content is disseminated.

5.3 Shared Recommendations:

The main recommendation is for both frameworks to recognize AI-generated non-consensual deepfakes as a form of gender-based violence, as the gender-neutral Acts obscure the fact that AI-enabled sexual abuse disproportionately affects women and girls. Formal recognition of GBV would ensure stronger legal protections, victim support, and reporting mechanisms specifically designed for this kind of harm. Similarly, the frameworks rely on universal categories and provide limited recognition of how harm affects different groups. According to feminist governance, it is necessary to recognize gender, race, sexuality, disability, age, and other categories in order to assess intersecting inequalities. Finally, cross-border enforcement is a significant challenge in the AIA and OSA, as regulatory responses remain within individual jurisdictions. The issue is that once AI-generated NCII is shared online, it can easily cross borders where the frameworks are not applicable. Therefore, international cooperation is needed to remove harmful content across borders and establish more effective reporting mechanisms.

6. Conclusion:

Artificial intelligence-generated non-consensual intimate images are representative of the rapid evolution of AI-enabled image-based abuse that poses significant challenges to the two examined frameworks. While both the EU AI Act and the UK Online Safety Act include necessary regulatory mechanisms, they do not address the gendered nature of harm, nor do they provide

comprehensive regulation of non-consensual deepfakes as a form of gender-based violence. Through a feminist governance perspective, both Acts need to recognize structural inequalities, center the experiences of victims, prioritize prevention, and distribute responsibility among the actors involved in the creation, deployment, and dissemination of AI systems. As generative AI continues to evolve, policymakers must ensure the protection of victims from emerging forms of digital gender-based violence.

7. **References:**

Akter, M.S. and Ahmed, P. (2025) 'The emergence of AI-generated deepfakes as a new tool for gender-based violence against women: A brief narrative review of evidence and the implications of the techno-feminist perspective', *Feminists@law*, 13(2). Available at: https://doi.org/10.22024/UniKent/03/fal.1281

Casagrande, R. and Spaticchia, U. (2025) 'New forms of gender-based violence: Between deepreal and deepfake', *Studia Humanitatis Journal*, 5. Available at: https://doi.org/10.33732/shj.v5i1.110

Conger, K., Freedman, D. and Thompson, S.A. (2026) 'Musk's chatbot flooded X with millions of sexualized images in days, new estimates show', *The New York Times*, 22 January. Available at: https://www.nytimes.com/2026/01/22/technology/grok-x-ai-elon-musk-deepfakes.html

European Commission (2026) *EU agrees to simplify AI rules to boost innovation and ban 'nudification' apps to protect citizens*. Brussels: European Commission. Available at: https://digital-strategy.ec.europa.eu/

Karagianni, A. and Doh, M. (2024) 'A feminist legal analysis of non-consensual sexualized deepfakes: Contextualizing its impact as AI-generated image-based violence under EU law', *Porn Studies*, 0(0). Available at: https://doi.org/10.1080/23268743.2024.2408277

Laffier, J. and Rehman, A. (2023) 'Deepfakes and harm to women', *Journal of Digital Life and Learning*, 3(1). Available at: https://doi.org/10.51357/jdll.v3i1.218

Belen Bringas
15/06/2026
AIPL Externship

Milmo, D. (2026) ‘Grok scandal highlights how AI industry is “too unconstrained”, tech pioneer says’, *The Guardian*, 15 January. Available at: https://www.theguardian.com/technology/2026/jan/15/grok-scandal-ai-industry-too-unconstrained-yoshua-bengio-elon-musk

Okolie, C. (2023) ‘Artificial intelligence-altered videos (deepfakes), image-based sexual abuse, and data privacy concerns’, *Journal of International Women’s Studies*, 25(2). Available at: https://vc.bridgew.edu/jiws/vol25/iss2/11

Prado, E.L.-A. (2025) ‘Reconceptualising the digital gender divide, accommodating new forms of virtual gender-based violence’, *Behavioral Sciences*, 15(11). Available at: https://doi.org/10.3390/bs15111568

Rigotti, C., McGlynn, C. and Benning, F. (2024) ‘Image-based sexual abuse and EU law’, *German Law Journal*, 25(9). Available at: https://doi.org/10.1017/glj.2024.49

Sawer, M., Banaszak, L.A., True, J. and Kantola, J. (2023) ‘Introduction to the handbook of feminist governance’, in Sawer, M., Banaszak, L.A., True, J. and Kantola, J. (eds.) *Handbook of Feminist Governance*. Edward Elgar Publishing.

Shade, L.R. and Shepherd, T. (2025) ‘Policy Portal 5: Intersectionality and Inclusion in Digital Policy: Intersectional Critiques of AI Governance’, *Canadian Journal of Communication*, 50(1). Available at: https://doi.org/10.3138/cjc-2024-0083

Williams, K. (2025) ‘“There are no limits!”: AI undressing apps and the normalization of nonconsensual intimate deepfakes’, *Violence Against Women*. Available at: https://doi.org/10.1177/10778012251397966