Pornography Regulation, Legislation and Enforcement
Response to the UK Government's Call for Evidence

Evidence submitted by Dr Beatriz Kira, Lecturer in Law at the University of Sussex and Fellow in Law and Regulation at UCL's Digital Speech Lab[1]

7 March 2024

Summary

1. To effectively counter **Non-Consensual Intimate Deepfakes (NCID)**, existing regulatory frameworks should be strengthened to recognise NCID as a form of image-based abuse, subject to the same rules that govern other forms of non-consensual intimate disclosure, while acknowledging the unique challenges posed by the ease of creation and dissemination of synthetic media.
    a. Crucially, legislation and policy should adopt the clear and unambiguous language of "non-consensual intimate deepfakes" that recognises NCID as a form of image-based abuse and clearly differentiates it from legal pornography. The term "deepfake pornography" should not be used to refer to NCID.

2. Current legal measures primarily focus on *individuals* who share or threaten to share NCID content, putting less emphasis on the roles *other actors* should play in preventing its further distribution, in particular social media platforms and pornography platforms. While the Online Safety Act 2023 (OSA) expands the scope of relevant criminal offences (likely facilitating the prosecution of individuals), it provides limited measures to prevent dissemination and empower victims once content is shared online.
    a. The OSA's "systems and processes" approach to illegal content relies on platforms' policies for content removal, but these policies are often ambiguous regarding the treatment of NCID (e.g. whether it is covered under existing prohibitions of nudity, or bullying, harassment, and abuse), leading to inconsistent and potentially ineffective enforcement. Platforms' policies applicable to image-based abuse should be media- and technology-neutral, applying symmetric treatment to both real and synthetic/manipulated content. Crucially, content removal should not hinge on identifying malicious intent, nor rely solely on victims' reporting.

    b. With the OSA set to overhaul the Video-Sharing Platforms (VSP) regime, meaning it will likely apply to a wider range of online pornography providers, there is an opportunity to strengthen the regime beyond its current focus on protecting *viewers*. This can be achieved by encouraging platforms to adopt measures that protect and empower *victims* of NCID. Such measures should be incorporated into the forthcoming Ofcom guidance on protecting women and girls, which is expected by Spring 2025.

3. Upcoming AI legislation should go beyond creating rules on risk assessment and risk mitigation during model development and deployment. For any generative AI tools accessible to the public, regulation should require the implementation of comprehensive and enforceable content moderation systems. These systems must explicitly prohibit the creation of intimate synthetic media, with consistent and effective enforcement of this prohibition.

[1] I am grateful to Sarah Fisher and Jeffrey Howard for helpful comments and suggestions, and to Luke Richards for research assistance. All errors remain my own. This evidence is based on the findings of the working paper Kira, B. "When Deepfakes Go Viral: Non-Consensual Intimate Deepfakes Under the UK Online Safety Act". Please contact me at b.kira@sussex.ac.uk for the most recent version of the manuscript.

1. What are the harms of non-consensual intimate deepfakes (NCID)?

Non-consensual intimate deepfakes (NCID), what is colloquially and mistakenly called "deepfake porn" or "deepfake pornography", should be in the scope of the UK government's review. The review should consider whether there are sufficient protections for NCID's victims in the context of reviewing regulation, legislation, and enforcement.

Deepfakes are a specific type of synthetic media. The term is made by the combination of the term "deep learning" and "fake" and used to describe the output created by using deep learning algorithms to synthetise a person's likeness in a video, image, or audio.[2] The Oxford English Dictionary defines deepfake as "Any of various media, esp. a video, that has been digitally manipulated to replace one person's likeness convincingly with that of another, often used maliciously to show someone doing something that he or she did not do."[3] Different computational techniques can be employed to produce deepfakes, but at the core of the concept is the idea of using AI tools to create media that represent or replace the likeness of a real person.[4]

While some deepfakes can be harmless or even beneficial,[5] they also pose significant risks. As AI tools for creating deepfakes become more accessible and better at producing realistic-looking videos and photos, significant attention has focused on the potential for fabricated media to cause harm, particularly in relation to misinformation and democratic processes.[6] However, a crucial – and often overlooked[7] – area of concern lies in the harm caused by non-consensual intimate deepfakes, a form of image-based sexual abuse.[8]

To be clear, the harms caused by the non-consensual sharing of intimate photos and messages predate AI, and have existed long before AI generated tools became widely available. In the UK alone, over 28,000 cases of non-consensual disclosure of private sexual images were reported to police between April 2015 and December 2021.[9] Like other forms of online harms, this disproportionately impacts women, leading to serious consequences for their well-being, including negative mental and physical health effects.[10]

---

[2] Ruben Tolosana and others, 'Deepfakes and beyond: A Survey of Face Manipulation and Fake Detection' (2020) 64 Information Fusion 131.

[3] 'Deepfake' <https://www.oed.com/dictionary/deepfake_n?tab=meaning_and_use#1345352340> accessed 3 March 2024.

[4] The concept of generative AI is contested, and generative AI tools can encompass applications that create different forms of content, including text, audio, images, videos, and even 3D models. Some popular examples include ChatGPT for text, Midjourney for images, and DeepBrain for videos. In the context of this brief, "AI tools" refers specifically to applications available to the public that can generate visually realistic images or videos based on textual prompts. For more details about the different applications of generative AI, see Fiona Fui-Hoon Nah and others, 'Generative AI and ChatGPT: Applications, Challenges, and AI-Human Collaboration' (2023) 25 Journal of Information Technology Case and Application Research 277.

[5] For example, research shows that using deepfakes to personalise training tools by replacing trainers' faces with users' improves learning outcomes. Christopher Clarke and others, 'FakeForward: Using Deepfake Technology for Feedforward Learning', *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (ACM 2023) <https://dl.acm.org/doi/10.1145/3544548.3581100> accessed 23 February 2024.

[6] The specific misinformation harms related to electoral deepfakes is discussed in a manuscript by Sarah Fisher, Jeffrey Howard and Beatriz Kira, "Moderating Misinformation: The Challenge of Generative AI" (manuscript).

[7] Research has documented the media's emphasis on misinformation harms associated with deepfakes. See Chandell Gosse and Jacquelyn Burkell, 'Politics and Porn: How News Media Characterizes Problems Presented by Deepfakes' (2020) 37 Critical Studies in Media Communication 497.

[8] Clare McGlynn and Erika Rackley, 'Image-Based Sexual Abuse' (2017) 37 Oxford Journal of Legal Studies 534.

[9] UK Government, 'Government Crackdown on Image-Based Abuse' (27 June 2023) <https://www.gov.uk/government/news/government-crackdown-on-image-based-abuse> accessed 3 March 2024.

[10] Online Violence Against Women Observatory, 'OU Research Reveals Shocking Level of Online Violence Experienced by Women and Girls across the UK' (13 September 2023) <https://www5.open.ac.uk/ovaw-observatory/news/ou-research-reveals-shocking-level-online-violence-experienced-women-and-girls-across-uk> accessed 3 March 2024.

While not novel, AI-generated deepfakes present new challenges. Unlike traditional non-consensual media which depends on the existence of real content, depicting actual events, NCID can be entirely fabricated. This significantly expands the potential volume of harmful material and makes anyone a potential target. Additionally, advancements in AI have made creating synthetic media easier and cheaper, potentially leading to a surge in this form of harassment and intimidation.

Therefore, it is crucial that any regulation and legislation that apply to non-consensual disclosure of real images explicitly encompass synthetic and manipulated media as well. Enforcement needs to account for the increased volume of illegal content and difficulty in preventing the spread of NCID due to the ease and speed with which AI tools generate this harmful content.

Crucially, legislation, regulation, and policy should use clear, specific language that accurately reflects the nature of the harms. Language used to describe content affects perceptions of the harms it causes. NCID should not be conflated with legal pornography. Terms like “deepfakes porn” or "AI-generated pornography" are misleading and risk conflating legal and illegal content. Assigning the label “pornography” to NCID trivialises the content and shifts the focus to its usage or interpretation, rather than the victim’s experience.[11] The term “non-consensual intimate deepfakes” accurately describes the illegal nature and victim impact, making it the most appropriate term for legislative, regulatory, and policy purposes.

2. Are the obligations introduced in the Online Safety Act 2023 sufficient to counter NCID?

Sharing NCID online significantly amplifies the harm it causes to victims. Both social media and pornography platforms enable the spread of harmful deepfakes, and researchers have long argued that online platforms amplify image-based sexual abuse harms, even before the rise of generative AI.[12] Through search results and algorithmic recommendations, they can act as digital firehoses of NCID, exponentially increasing the reach of the content beyond the initial perpetrator and potentially exposing the victim to a vast, unknown audience. This wider exposure aggravates the emotional distress, rights’ violation, and potential safety threats faced by victims.

The recently introduced Online Safety Act 2023 (OSA) is the key legislation governing online platforms in the UK. Its rules are also set to replace the current Video-Sharing Platforms (VSP) regime, which applies to many pornography platforms. Therefore, the OSA is poised to become the main law regulating online pornography, with Ofcom the primary regulator. While the OSA strengthens criminal law by expanding the scope of previous provisions and simplifying the prosecution of individuals, the obligations it places on platforms, Ofcom’s illegal content judgement guidance, and platforms’ content policies offer limited remedies to victims of NCID once the content is already online. I examine each of these aspects of the OSA regime below.

[11] Professor Clare McGlynn and Professor Erika Rackley have convincingly argued against using the term “revenge pornography” to describe non-consensual sharing of intimate content, advocating for the more accurate term “image-based sexual abuses”. Their arguments extend directly to the content of AI-generated image-based abuses. See, for example, Clare McGlynn and Erika Rackley, ‘Not “Revenge Porn”, but Abuse: Let’s Call It Image-Based Sexual Abuse’ (*Inherently Human*, 15 February 2016) <https://inherentlyhuman.wordpress.com/2016/02/15/not-revenge-porn-but-abuse-lets-call-it-image-based-sexual-abuse/> accessed 2 March 2024; McGlynn and Rackley (n 8). For a discussion of how “revenge porn” was reframed and renamed, see also Sophie Maddocks, ‘From Non-Consensual Pornography to Image-Based Sexual Abuse: Charting the Course of a Problem with Many Names’ (2018) 33 Australian Feminist Studies 345.

[12] McGlynn and Rackley (n 8); Nicola Henry and Anastasia Powell, ‘Sexual Violence in the Digital Age: The Scope and Limits of Criminal Law’ (2016) 25 Social & Legal Studies 397.

2.1. Criminal offences

The OSA strengthens protections for victims of image-based abuse by introducing new offences related to sharing or threatening to share intimate pictures or videos (adding new sections to the Sexual Offences Act). This broader approach aims to offer more robust protection to victims by eliminating the need for specific criminal intent (*mens rea*) in the base offence, potentially facilitating prosecution.[13] The new language also seems to encompass synthetic content like deepfakes. Two important changes are noteworthy. First, the offences reference photography or film that "shows or appears to show" someone, which means it could apply to content that appears real despite being completely fabricated. Second, the new wording broadens the definitions of "photography" and "film" to explicitly include images "made or altered by computer graphics or in any other way, which appears to be a photograph or film" (Section 66A(5) Sexual Offences Act). Therefore, to avoid any confusion, any guidance to law enforcement should clarify that these offences apply equally to both real and synthetic non-consensual content.

2.2. User-to-user services' duties

In addition to the new criminal offences, the OSA seeks to strengthen protection of victims by placing "duties of care" on user-to-user internet services (including social media platforms), requiring them to implement safety measures to curb illegal content and content harmful to children. Notably, sharing non-consensual intimate photographs and films fall under the priority offence category (Schedule 7, para. 30, OSA), triggering the stricter duties associated with this category.[14]

For priority offences, the Act requires platforms to use "proportionate measures" in their design and operation to prevent users from encountering priority illegal content and mitigate the risk of the service being used for related offences (Section 10(2) OSA). Notably, platforms are not required to prevent every encounter between a user and priority illegal content, but to adopt proportionate measures to minimise such encounters. Platforms' duty is to reduce the presence of this content, not to guarantee that none of it is ever available.

This reflects the OSA's systemic approach, placing the onus on platforms to police content. Ofcom's guidance clarifies that platform compliance will not be judged based on the absence of individual illegal content pieces but on their systematic efforts to mitigate risks. The guidance further recognises that services are likely to deal with "content in bulk" rather than individually and that platforms can address illegal content on a "probabilistic basis", meaning they are not required to identify and remove every single instance of illegal content.[15] Clear guidance from Ofcom on assessing compliance with these requirements at a systemic level is lacking.[16]

---

[13] Several scholars have noted the limitations of previous legislation, notably section 33 of the Criminal Justice and Courts Act (CJCA) 2015. See Alisdair Gillespie, '"Trust Me, It's Only for Me": Revenge Porn and the Criminal Law' [2015] Criminal Law Review 866; McGlynn and Rackley (n 8); Henry and Powell (n 12).

[14] There is a potential discrepancy within the Online Safety Act. Schedule 7, paragraph 30, references "An offence under section 33 of the Criminal Justice and Courts Act 2015 (disclosing, or threatening to disclose, private sexual photographs and films with intent to cause distress)" as a priority offence. However, sections 33 to 35 of the aforementioned Act are repealed by section 190 of the OSA due to the introduction of new sexual offences in the Sexual Offences Act. As a result, the logical interpretation suggests that Schedule 7 refers to the newly established sexual offences as the priority illegal content, not the repealed ones.

[15] Ofcom, 'Annex 10: Online Safety Guidance on Judgement for Illegal Content' (2023) para A1.15 <https://www.ofcom.org.uk/__data/assets/pdf_file/0025/271168/annex-10-illegal-harms-consultation.pdf> accessed 29 February 2024.

[16] The shortcomings of Ofcom's guidance on illegal content assessment and implications for freedom of expression are developed in more detail in Ellen Judson, Beatriz Kira, and Jeffrey W. Howard, "Bypassing Legal Judgement: Platforms, the Online Safety Act, and Future of Online Speech" (manuscript).

Ofcom's draft illegal content judgment guidance focuses heavily on interpreting individual pieces, suggesting that platforms should "look for identifier features which suggest lack of consent and either lack of reasonable belief in consent or intent to cause alarm, distress or humiliation".[17] To make such assessment, platforms should be informed by users reporting lack of consent, contextual information indicating the malicious intent, or contextual information indicating the media has been "leaked" or "hacked". Ofcom's guidance clearly misses the mark on how platforms should deal with non-consensual sharing of content, and is particularly problematic for NCID, for three main reasons.

First, the guidance relies heavily on identifying "signs of non-consent" within the content itself for removal. However, a significant portion of non-consensual content is shared without explicit mentions of hacking or lack of consent, and often lacks signs of any malicious intent. Even with real explicit content, the victim might have initially consented to the recording, but not its sharing – therefore the footage itself will not reveal this lack of consent. Furthermore, tools used to create NCID can easily manipulate content, erasing any trace of non-consent that could have been present originally. Therefore, while signs of non-consent should certainly lead platforms to remove content, these signs should not be a requirement for content removal, as this aspect will often be lacking.

Secondly, and crucially, the guidance relies heavily on victim reporting, despite experts warning that many victims of online violence do not report.[18] In addition, when it comes to fully synthetic content, victims will be completely unaware of the material's existence since they were not involved in its creation, further hindering their ability to report it.

Thirdly, the guidance only applies for *reactive* forms of content moderation by platforms, after the content has already been online for enough time for someone to flag it as non-consensual, when harm will already have been caused. However, identifying consent requires contextual understanding, which is beyond the capabilities of automated moderation systems. This limitation hinders *proactive* measures that could be far more effective in preventing harm, such as content moderation methods that leverage technology to detect and stop the spread of synthetic intimate material.

In reality, the identification of the mental elements that are relevant from a law enforcement perspective to identify which of the specific offences in Section 66B of the Sexual Offences Act have been committed is completely irrelevant from the perspective of content moderation by platforms. If any form of intimate content is posted or shared without the consent of the individual being portrayed, platforms should take action. Ofcom's responsibility is to provide clear steer on this point, which their current guidance fails to do, ultimately undermining the effectiveness of the OSA's safety duties.

2.3. Platforms' policies

Given the OSA's systems and processes approach, the effectiveness of remedies for victims depends heavily on the rules and processes adopted by each platform. However, as the table below demonstrates, these policies offer varying levels of protection for non-consensual intimate synthetic content, or NCID:

[17] Ofcom (n 15) A10.31.

[18] A coalition of 44 specialist organisations and experts in violence against women and girls (VAWG) has expressed several concerns with the Ofcom's approach to tackling illegal online content, including the fact that the approach is places too much focus on reporting as an indicator of harm, when in fact many victims don't report. 'Open Letter - Ofcom Blocking a Safer Internet for Women, VAWG Experts Warn' (23 February 2024) <https://www.endviolenceagainstwomen.org.uk/ofcom-blocking-a-safer-internet-for-women-vawg-experts-warn/> accessed 4 March 2024.

Table 1. Do platforms' main content policies apply to non-consensual intimate deepfakes?

	Meta	TikTok	X (formerly Twitter)
Nudity	No – rule applies to real nude adults	Ambiguous – unclear whether rule applies to synthetic media	Yes – "non-consensual nudity" rule clearly applies to deepfakes
CSAM	Yes – clearly applies to both real and synthetic content	Yes – clearly applies to both real and synthetic content	Yes – clearly applies to both real and synthetic content
Bullying, harassment and abuse	Ambiguous – unclear whether rule applies to synthetic media	Ambiguous – unclear whether rule applies to synthetic media	Ambiguous – unclear whether rule applies to synthetic media
Synthetic media	No – although Meta has a policy prohibiting synthetic videos where "a subject of the video said words that they did not say"	Yes – policy explicitly prohibits synthetic media that "contains the likeness of any real private figure"	Yes – synthetic media "may not" be shared if it "may deceive or confuse people and lead to harm", however, the sanction could simply be that it is labelled as misleading by X

Source: Kira, B. "When Deepfakes Go Viral: Non-Consensual Intimate Deepfakes Under the UK Online Safety Act".[19]

Social media platforms currently have an inconsistent approach to NCID, as evidenced by the table. Unlike their clear and consistent policies towards child sexual abuse material (CSAM), which encompass artificial content, current adult content policies are ambiguous, potentially creating loopholes and hindering the protection of victims. To effectively counter NCID, clear symmetry is required. Platforms should implement robust rules prohibiting all image-based abuse, regardless of its origin or creation technology. Additionally, strong and accessible reporting mechanisms and quick response by platforms are crucial for prompt removal of illegal content. This is especially important as AI tools are making it easier and faster to create NCID, meaning platforms need to assess whether existing systems are adequate to keep up with the increase in the volume of non-consensual intimate content.

Ofcom has a critical role in preparing platforms to counter NCID more effectively under the Online Safety Act, especially since the act will replace the Video-Sharing Platforms (VSP) regime and apply to a wider range of online pornography services. As the VSP regime gets overhauled, the Online Safety Act provides an opportunity to offer better protection for victims. The VSP regime's obligations regarding "relevant harmful material" and "restricted material" primarily centred on protecting viewers from harmful types of content, rather than the harm content could inflict on individuals depicted in the content. The Online Safety Act's overhaul of this regime presents a renewed opportunity to strengthen the focus on *victims*, encompassing both adult content providers and individuals whose likeness is exploited in NCID. Given the disproportionate impact on women, Ofcom's guidance on protecting women and girls, expected by Spring 2025, will be crucial. This guidance should clearly communicate platforms' obligation to implement robust protections to victims of NCID. This could involve, for example, clarifying interpretations of illegal content duties related to relevant offenses and suggesting that platforms establish expedited procedures for processing complaints.[20]

[19] Working paper. Please contact the author for the most recent version of the manuscript.

[20] Ofcom, 'How the Online Safety Act Will Help to Protect Women and Girls' (29 November 2023) <https://www.ofcom.org.uk/news-centre/2023/how-the-online-safety-act-will-help-to-protect-women-and-girls> accessed 3 March 2024.

3. How can forthcoming AI regulation strengthen the regulatory regime?

The user-friendliness and online accessibility of AI tools have significantly facilitated the creation of NCID. While pre-existing tools could be used for malicious purposes, AI tools have significantly lowered the entry barrier for bad actors. As cheap, easy-to-use online tools that generate increasingly convincing victim likenesses, become more available, they contribute to the surge in NCID.

Given the ease with which these AI tools can be misused, it is crucial to establish safeguards to prevent the creation of NCID. This requires including AI tools, specifically those with a user interface and the capability to generate synthetic intimate content, to put in place appropriate safeguards. However, under the current definitions, generative AI tools likely fall outside the scope of the Online Safety Act because they do not meet the criteria of user-to-user services or search engines. This raises the question of whether additional regulations are needed to hold both the developers and distributors of these tools accountable for ensuring that they are not used to create or distribute NCID.

The "pro-innovation approach" to AI regulation proposed by the UK government, while positive in encouraging development of new technologies, currently lacks tools to effectively address NCID. The government's report, when discussing deepfakes, primarily focuses on "AI-related risks to trust in information".[21] The only mention of intimate image abuse addresses how existing issues are handled under the Online Safety Act, which, as discussed above, offers limited protection in this specific context.

Forthcoming AI regulations can incentivise developers of generative AI and those who make these tools available to the public to actively reduce the risk of harm by minimising harmful content generated by their models from the outset. This is crucial, as current usage policies of some of the main tools (including OpenAI and Gemini) are minimal compared to social media platforms' content policies. OpenAI, for example, merely asks users to avoid compromising "the privacy of others", engaging in "regulated activity without complying with applicable regulations", or promoting/engaging in "any illegal activity". In addition, there are no details on how such policies are enforced. Vague policies are unlikely to offer robust protections.

To address the gap in safeguards against non-consensual intimate deepfakes, AI regulation should require developers and distributors of generative AI tools to implement robust safety measures, mirroring the private systems of governance used by major social media companies to regulate content on their platforms.[22] Specifically, AI regulation should legally mandate AI tools to adopt clear usage policies prohibiting the creation of synthetic intimate content depicting real people. Additionally, regulation should require AI tools to have effective enforcement measures in place to ensure consistent application of the established policies.

[21] UK government, 'A Pro-Innovation Approach to AI Regulation: Government Response' (2024) Unique Reference: E03019481 02/24 para 80 <https://www.gov.uk/government/consultations/ai-regulation-a-pro-innovation-approach-policy-proposals/outcome/a-pro-innovation-approach-to-ai-regulation-government-response> accessed 3 March 2024.

[22] On the lessons content moderation on social media platform can offer to prevent and mitigate harms of generative AI, see Jeffrey Howard and Beatriz Kira, 'How Should We Regulate LLM Chatbots? Lessons from Content Moderation' (*The Digital Constitutionalist*, October 2023) <https://digi-con.org/how-should-we-regulate-llm-chatbots-lessons-from-content-moderation/> accessed 30 October 2023.