Deepfakes of Intimate Images

Introduction

In January 2024, a disturbing incident unfolded online: sexually explicit deepfake images of Taylor Swift [12] spread like wildfire across social media platforms, racking up millions of views. These images, created using artificial intelligence, depicted Swift at a football game, engaging in a scenario that never occurred. Within hours, these images had reached over 45 million views, sparking outrage and concern. Despite clear violations of platform policies, the images remained accessible for an alarming amount of time, leaving Swift and her millions of followers to grapple with the personal toll of such a violation. This incident is not unique to Swift or limited to public figures.

For instance, 11 months ago, a New Jersey high school sophomore named Francesca Mani found herself at the center of a similar nightmare. Over the summer, a group of boys at her school used artificial intelligence to create sexually explicit images of several girls, including Francesca[13]. These deepfakes circulated among peers, subjecting the victims to harassment, humiliation, and emotional distress.

These two cases, one involving a global star and the other a high school student, highlight the terrifying accessibility and reach of deepfake technology. They show how anyone, regardless of age or fame, can become a target of AI-generated sexual exploitation. The emotional and reputational damage can be profound. Yet, current systems often fail to respond swiftly or adequately. These failures raise critical questions: Who is responsible for the spread of such content? What legal protections exist for victims? How can the laws be updated to keep pace with the rapid evolution of this technology?

Legal Background

The legal framework surrounding deepfake technology remains incomplete, particularly regarding AI-generated non-consensual intimate images (NCII). Deepfake technology is a relatively new and rapidly evolving area of digital technology, which complicates the creation of laws that are both effective and adaptable to the fast pace of innovation. Despite that, many laws exist to attempt to protect the everyday person from NCII.

The first major step was the PROTECT Act [1], officially known as the PROTECT Act of 2003: Prosecutorial Remedies and Other Tools to End the Exploitation of Children Today Act. While the primary focus of the law was on the possession and distribution of Child Sexual Abuse Material (CSAM), the act also introduced provisions that could be applied to the manipulation of images. This includes the use of digital technology to create or alter exploitative content. Essentially, the PROTECT Act laid the groundwork for addressing digital forms of exploitation, but its application to AI-generated deepfakes remains limited. [1]

Then, there is U.S. Code 18, Section 2252 [2], which criminalizes the possession,

distribution, and production of CSAM, including some forms of digitally altered content. However, like the PROTECT Act, its scope is primarily focused on content involving victims of NCII. This leaves a significant legal gap for AI-generated CSAM and NCII, as the law does not explicitly address synthetic content or digital manipulation that does not involve actual children. In addition, while Section 2252 of the U.S. Code 18 criminalizes the production, possession, and distribution of digitally altered CSAM, it does not sufficiently address the technological advancements that enable the creation of realistic, synthetic, exploitative images and videos. This leaves a loophole where AI-generated content, which may not involve real children or even real people, remains unregulated. As a result, bad actors can produce and share deepfakes of children or adults in exploitative scenarios, creating a new class of harm that falls outside the scope of existing laws. [2]

Another law tackling this issue is Section 230 [3] of the Communications Decency Act (CDA), a key legal framework in the regulation of online content. It has been crucial in the development and growth of the internet, but it also presents significant challenges in addressing the proliferation of harmful digital content, including deepfakes. As the law states, "No interactive computer service provider or user shall be treated as the publisher or speaker of any information provided by another information content provider." [3] This section effectively grants online platforms immunity from being held liable for user-uploaded content; it allows companies like social media networks, video-sharing platforms, and forums to function as intermediaries without assuming legal responsibility for the vast amount of content generated by users. However, while fostering a free and open internet, this legal immunity has come under increasing scrutiny, especially in cases involving harmful content such as deepfakes. The issue with Section 230 in the context of AI-generated NCII and CSAM, lies in its broad application. Because platforms are not treated as publishers or speakers of user-generated content, they are not required to take proactive steps in policing the content posted on their sites. As a result, platforms have little legal incentive to remove malicious content, such as deepfakes, unless it explicitly violates their terms of service. This creates a regulatory blind spot, allowing harmful content to circulate widely without significant accountability. In the case of deepfake technology, the consequences of Section 230 are particularly evident. Deepfakes, by their nature, involve the manipulation of real and/or synthetic media to create highly convincing yet entirely fabricated representations of individuals, often with explicit or exploitative intent. This lack of accountability does nothing to help victims [3]

Technical Background

Deepfakes are hyper-realistic videos, images, or audio recordings generated using artificial intelligence (AI) algorithms. They involve digitally altering content to depict individuals performing actions or speaking words they never did. This is achieved by aligning the faces of two individuals: an autoencoder captures features from "face A". It merges them with "face B", resulting in a synthetic depiction resembling B but not authentically representing their actual appearance. [14]

The creation of deepfakes primarily relies on deep neural networks, specifically autoencoders. An autoencoder consists of two main components: an encoder and a decoder. The encoder processes input images or videos, compressing them into a latent code that

retains essential features while discarding extraneous details. The decoder then reconstructs the original content from this latent representation. In the context of deepfakes, the encoder captures the facial features of the source individual ("face A"), and the decoder reconstructs these features in the context of the target individual ("face B"). This process results in a synthetic depiction that combines elements of both individuals. [14]

Another advanced technique employed in deepfake creation involves Generative Adversarial Networks (GANs). GANs consist of two neural networks: the generator and the discriminator, engaged in a continuous feedback loop. The generator creates synthetic content, while the discriminator evaluates its authenticity. Through this adversarial process, the generator improves its ability to produce highly realistic media, as it learns to create content that can deceive the discriminator. [14]

An innovative idea regarding GANs was first introduced in 2023 by Nadimpalli and Rattani. They focus on GAN-based visible watermarking to enhance the detection and verification of deepfake media [15]. The approach utilizes GANs with a visible watermark embedded in the generated media. The watermark is seamlessly integrated into the synthetic media during the generation process. The generator network, which is responsible for producing the deepfake content, is modified to include a watermark as part of its output. This watermark is not simply superimposed onto the generated content after its creation; instead, it is embedded within the content itself at the time of its creation through a specific modification in the generator's architecture. The watermark is designed to be imperceptible to the human eye, meaning it does not noticeably alter the visual quality of the deepfake content. However, the watermark is detectable through advanced analysis methods, ensuring that it can be used for verification purposes in the future [15].

The testing of the GAN-based visible watermarking approach was conducted to evaluate its effectiveness in detecting and verifying deepfake media while maintaining the realism of the generated content. The researchers employed a comprehensive set of experiments to evaluate the robustness and imperceptibility of the watermark in various scenarios, including different attack methods designed to remove or distort the watermark.

The first test is the Imperceptibility test, which evaluates how noticeable the watermark is in the generated DeepFake images. The goal is to ensure that the watermark is barely visible to the human eye while still being detectable by the system for proactive detection. In this test, the Structural Similarity Index Measure (SSIM) and Fréchet Inception Distance (FID) were used to compare the quality of watermarked images to non-watermarked ones. The results showed a minimal reduction in image quality (SSIM = 0.48 for non-watermarked images and SSIM = 0.45 for watermarked ones, FID increased from 11.56 to 12.78), meaning that the watermarking has a very slight effect on image fidelity[15].

Next is the robustness test, which measures how well the watermark survives various attacks or image transformations. This test is crucial because it ensures that the watermark is resilient to methods like cropping, resizing, or other alterations that could be used to remove it. For example, a cropping attack (removing a portion of the image containing the watermark) resulted in a significant decrease in detection performance, with the AUC (Area Under the Curve) dropping from 0.902 to 0.789, indicating that cropping can reduce the watermark's effectiveness. This suggests that the watermark is vulnerable to specific attacks and may require additional refinement for enhanced resilience [15].

Then, the Discriminator Accuracy test in the paper evaluates how well the discriminator (a part of the GAN model) can differentiate between authentic images, fake images, and watermarked fake images. The discriminator is trained to distinguish between real and fake images, and additionally, it needs to detect the watermark embedded in the fake images. The results showed that the discriminator maintained high accuracy in distinguishing between fake images and real ones, even when the images were watermarked. The Area Under the Curve (AUC) for detecting fake images was 0.902, and for detecting watermarked counterfeit images, it was slightly lower at 0.897, indicating a minor drop in performance. This drop of 0.005 in AUC suggests that the watermark did not significantly impact the discriminator's ability to detect fake images. In essence, the discriminator could still effectively detect watermarked images, confirming that the watermark did not significantly interfere with the model's ability to distinguish between real and fake content, thereby maintaining the overall detection accuracy [15].

Lastly, the Adversarial Attack tests were used to evaluate deliberate attempts to remove or alter the watermark in generated images, thereby evading detection. Common adversarial attacks include fine-tuning (adjusting the GAN model to remove the watermark) and cropping (removing parts of the image containing the watermark). The results showed that fine-tuning had minimal impact on the watermark's detectability, but cropping caused a significant reduction in detection performance. The AUC dropped from 0.902 to 0.789, showing that cropping attacks are effective at disrupting the watermark. These findings suggest that while the watermarking technique is relatively robust, it is vulnerable to specific manipulations [15].

Proposal

The existing solutions and protections simply are not enough. My proposed solution consists of four main parts: criminalizing non-consensual deepfakes, enforcing platform responsibilities, ensuring liability for AI models, and promoting global cooperation.

While the Preventing Deepfakes of Intimate Images Act (H.R. 3106) [4] was introduced in 2023, it failed to advance in Congress, leaving a critical gap in addressing AI-generated non-consensual intimate images (NCII). The bill sought to criminalize the creation and distribution of non-consensual deepfake intimate images, but it did not go far enough in covering possession or addressing AI-generated CSAM [4]. Current laws, including the PROTECT Act and U.S. Code 18, Section 2252, only criminalize real CSAM and do not explicitly account for artificially generated media, meaning AI-generated CSAM (as well as AI-generated NCII) often falls outside legal definitions of criminal conduct [1][2]. To close this loophole, legislation should be updated to explicitly criminalize the creation, distribution, and possession of AI-generated NCII and CSAM; the use of AI tools to manipulate existing media to depict real individuals in explicit content without consent; and the intentional misuse of deepfake technology for sexual exploitation, blackmail, or reputational harm. By explicitly including AI-generated content in criminal statutes, the law would prevent bad actors from exploiting the existing legal gray areas.

The next component is enforcing platform responsibilities, and current laws, including Section 230 [3] provide online platforms with broad immunity from liability for user-generated content. Currently, legislation has been introduced that could establish this. The EARN IT Act (Eliminating Abusive and Rampant Neglect of Interactive Technologies Act of 2022) [5] seeks to

hold online platforms accountable for hosting CSAM by conditioning their Section 230 immunity on compliance with new child protection standards. This bill would establish a national commission responsible for developing best practices for detecting and removing CSAM from online platforms. [5] However, while the EARN IT Act targets CSAM specifically, it does not fully address AI-generated NCII. To improve upon the framework provided by the EARN IT Act, new legislation should expand liability to ensure that online platforms are responsible not only for real CSAM but also for AI-generated CSAM and deepfake NCII. Additionally, it should mandate proactive content moderation, requiring platforms to implement AI detection tools capable of identifying deepfake NCII before it spreads. A rapid takedown requirement, similar to the GDPR's "right to be forgotten", should also be introduced, allowing victims to request the immediate removal of manipulated content [6]. Additionally, it should mandate proactive content moderation, requiring platforms to implement AI detection tools capable of identifying deepfake NCII before it spreads. A rapid takedown requirement, similar to the GDPR's "right to be forgotten", should also be introduced, allowing victims to request the immediate removal of manipulated content [6]. Given the potential burden on smaller platforms, government subsidies or grants could be offered to help them implement these systems, along with a shared detection infrastructure, such as a national deepfake detection API maintained by a public-private partnership.

Then, there is the issue of holding AI models responsible for their role in enabling the creation of AI-generated NCII and CSAM. While current legislation focuses on the end users who create and distribute deepfake content, it largely ignores the responsibility of the companies developing and training AI models that can be exploited for this purpose. AI developers must be held accountable for ensuring that their models do not facilitate the generation of harmful content. A significant step toward this goal is the work done by the National Institute of Standards and Technology (NIST) AI Risk Management Framework (AI RMF), which provides guidelines for developing and deploying trustworthy AI systems. The NIST AI RMF emphasizes the need for AI models to be safe, secure, and accountable in their design and implementation [7]. However, these guidelines are voluntary, meaning that compliance is not legally enforced. To address this, legislation should mandate that AI developers incorporate safeguards preventing their models from being used to create explicit, non-consensual content. This includes restricting model training on pornographic material and embedding watermarking or traceability mechanisms to identify AI-generated content [7]. Furthermore, AI developers should be required to conduct regular audits and impact assessments, following the principles outlined in NIST's AI Trustworthiness Framework [8], to evaluate whether their models are being misused for harmful purposes. These assessments should include bias detection, risk evaluation, and proactive mitigation measures to prevent the weaponization of deepfake technology.

Similarly, a federal oversight agency or advisory board could be established to review AI systems for their potential for abuse. Developers who demonstrate due diligence and implement safeguards should be granted safe harbor protections, while those who fail to comply could be penalized or restricted from distributing their models. This approach would strike a balance between innovation and accountability, ensuring that both open-source and commercial AI tools are not misused.

Moreover, as mentioned earlier, a preventive method that shows promise is GAN-based visible watermarking, which could be mandated as a standard safeguard in generative AI systems. By embedding imperceptible but detectable watermarks directly into deepfake content during generation, this technique enables reliable attribution and facilitates detection without significantly compromising visual quality [15]. The research by Nadimpalli and Rattani (2023) demonstrated that such watermarking has a minimal impact on image fidelity while maintaining high detection accuracy, making it a viable solution for traceability. Although the approach showed some vulnerability to cropping attacks, its robustness against other manipulations and adversarial attacks suggests that, with further refinement, it could become a foundational element of responsible AI deployment with further refinement. Mandating the integration of visible watermarks into AI-generated content, especially content depicting human likeness, would enhance accountability, aid in rapid identification and removal of illicit deepfakes, and deter misuse by increasing the likelihood of detection and traceability. However, implementing GAN-based visible watermarking comes with significant trade-offs. As noted by Nadimpalli and Rattani, “The limitation of our proposed approach could be training GANs with the necessary regularization for watermark embedding could be resource-intensive and time-consuming” [15]. This means that incorporating watermarking mechanisms into GANs requires additional computational resources and extended training times, which may pose practical challenges for developers, especially those working with limited infrastructure. Moreover, the process of fine-tuning generator architectures to embed imperceptible watermarks without degrading image quality adds complexity to the model design. While these trade-offs do not undermine the value of visible watermarking, they underscore the need to strike a balance between security and performance. Future developments must focus on optimizing these techniques to reduce the computational burden while preserving their effectiveness in deepfake detection and traceability.

Lastly, the issues arising with deepfake technology are not limited to one state, country, or continent; it is a global problem. Such a global issue requires international collaboration. One model for such international collaboration is the General Data Protection Regulation (GDPR) [9], which governs data protection and privacy across the European Union (EU). Article 50 of the GDPR, titled "International Cooperation and Consistency," [10] highlights the importance of global coordination in enforcing data protection standards. It specifies that the European Data Protection Board (EDPB) must cooperate with non-EU countries and organizations to ensure consistent application of the regulation across borders. The language in Article 50 is particularly relevant when thinking about the regulation of AI-generated content, as it lays the groundwork for collaborative enforcement. However, in practice, international cooperation remains a complex issue due to differences in legal systems, cultural values, and regulatory frameworks across countries. According to a study, the extraterritorial reach of the GDPR has been met with resistance from several non-EU countries, which argue that the regulation imposes undue burdens on their legal systems and sovereignty. This resistance has led to difficulties in enforcing consistent global standards for data protection, as many countries have not fully adopted or aligned their own policies with those of the GDPR [11]. It is very likely that similar hardships might arise when attempting to implement global protection against AI-generated NCII and similar material.

By implementing these measures, we can create a legal and technological framework that not only protects individuals from the harm caused by non-consensual deepfakes but also sets a precedent for addressing emerging AI-driven threats.

Citations

1) S.151 - 108th Congress (2003-2004): Protect Act | Congress.Gov | Library of Congress, www.congress.gov/bill/108th-congress/senate-bill/151. Accessed 18 Mar. 2025.
2) "18 U.S. Code § 2252 - Certain Activities Relating to Material Involving the Sexual Exploitation of Minors." Legal Information Institute, Legal Information Institute, www.law.cornell.edu/uscode/text/18/2252. Accessed 18 Mar. 2025.
3) Product Details R46751 - CRS Reports - Congress.Gov, crsreports.congress.gov/product/details?prodcode=R46751. Accessed 18 Mar. 2025.
4) "Preventing Deepfakes of Intimate Images Act (2023 - H.R. 3106)." GovTrack.Us, www.govtrack.us/congress/bills/118/hr3106. Accessed 18 Mar. 2025.
5) Text - H.R.454 - 118th Congress (2023-2024): Preventing Child Sex Abuse Act of 2023 | Congress.Gov | Library of Congress, www.congress.gov/bill/118th-congress/house-bill/454/text. Accessed 18 Mar. 2025.
6) Keller, Daphne. "The GDPR's Notice and Takedown Rules: Bad News for Free Expression, but Not beyond Repair." Stanford CIS, Stanford CIS, 29 Oct. 2015, cyberlaw.stanford.edu/blog/2015/10/gdprs-notice-and-takedown-rules-bad-news-free-expression-not-beyond-repair/.
7) NIST, nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf. Accessed 18 Mar. 2025.
8) "Trustworthy and Responsible AI." NIST, 4 May 2023, www.nist.gov/trustworthy-and-responsible-ai.
9) "GDPR." General Data Protection Regulation (GDPR), 22 Apr. 2024, gdpr-info.eu/
10) "Art. 50 GDPR – International Cooperation for the Protection of Personal Data." General Data Protection Regulation (GDPR), 6 Oct. 2016, gdpr-info.eu/art-50-gdpr/
11) Saqr, Mohammed. "Is GDPR Failing? A Tale of the Many Challenges in Interpretations, Applications, and Enforcement." International Journal of Health Sciences, U.S. National Library of Medicine, 2022, pmc.ncbi.nlm.nih.gov/articles/PMC9441646/.
12) "Taylor Swift and the Dangers of Deepfake Pornography." *National Sexual Violence Resource Center*, www.nsvrc.org/blogs/feminism/taylor-swift-and-dangers-deepfake-pornography Accessed 7 Apr. 2025.
13) Ryan-Mosley, Tate. "A High School's Deepfake Porn Scandal Is Pushing US Lawmakers into Action." *MIT Technology Review*, MIT Technology Review, 1 Dec. 2023, www.technologyreview.com/2023/12/01/1084164/deepfake-porn-scandal-pushing-us-lawmakers/.
14) Alanazi, Sami, and Seemal Asif. "Exploring Deepfake Technology: Creation, Consequences and Countermeasures - Human-Intelligent Systems Integration." *SpringerLink*, Springer International Publishing, 18 Sept. 2024, link.springer.com/article/10.1007/s42454-024-00054-8#:~:text=Deepfakes%20are%20produced%20using%20deep,(Juefei%2DXu%20et%20al.

15) Nadimpalli, Aakash Varma, and Ajita Rattani. *Proactive Deepfake Detection Using Gan-Based Visible Watermarking | ACM Transactions on Multimedia Computing, Communications, and Applications*, dl.acm.org/doi/full/10.1145/3625547. Accessed 8 Apr. 2025.