“Unwilling Avatars” revisited: A technical, legal, and social analysis of AI-generated nonconsensual intimate imagery

ABBY ROCHMAN

FAMILY ONLINE SAFETY INSTITUTE
ONLINE SAFETY FELLOWS PROGRAM
MAY 2025

Abby Rochman

"Unwilling Avatars"[1] revisited: A technical, legal, and social analysis of AI-generated nonconsensual intimate imagery

Executive Summary

This white paper examines the technical, social, and legal dimensions of AI-generated nonconsensual intimate imagery (NCII), a form of image-based sexual abuse facilitated by advancements in generative AI. It explores the mechanisms behind these systems to explain how tools like convolutional neural networks (CNNs) and generative adversarial networks (GANs) enable the creation of manipulated content. AI-generated NCII has profound consequences: victim-survivors often experience trauma, reputational damage, social isolation, and professional consequences. When minors are involved, the imagery may legally constitute child sexual abuse material (CSAM), which raises serious legal and ethical issues. Despite these harms, current laws and content moderation practices struggle to address the scale and speed of the threat.

This paper outlines three core dimensions of the problem:

- Technical: How AI nudification tools function, the role of open-source models, and emerging technical mitigation strategies such as detection, hashing, and reinforcement learning.
- Social: How gender norms, online anonymity, and the normalization of abuse inform the creation and spread of NCII.

[1] Mary Anne Franks, "Unwilling Avatars: Idealism and Discrimination in Cyberspace," *Columbia Journal of Gender and Law* 20 (2011)

- Legal and policy: The fragmented and inconsistent landscape of laws addressing deepfakes, and a review of promising legislation such as the TAKE IT DOWN Act and the DEFIANCE Act.

The report also spotlights promising industry and civil society responses, such as Microsoft's PhotoDNA and StopNCII.org's takedown tools. However, it argues that technical tools alone are not sufficient. This is an opportunity to shift culturally–to prioritize comprehensive sexual education, digital literacy, and product design choices that discourage exploitation.

Key recommendations include:

- Establish laws to criminalize AI-generated CSAM, and require platforms to implement faster, cross-platform takedown mechanisms of NCII and AI-generated CSAM, such as the TAKE IT DOWN Act and the DEFIANCE Act.
- Strengthen transparency and responsibility among AI developers, especially those releasing open-source tools.
- Integrate child and user safety principles throughout the AI development lifecycle.
- Invest in educational programs that teach youth about digital ethics, consent, and the consequences of image-based abuse.

By combining legal, technical, and cultural strategies, policymakers, technologists, and educators can help build an online ecosystem that respects privacy, protects autonomy, and minimizes the misuse of generative AI tools.

Introduction

While the recent AI boom has sparked optimism about the speed of innovation, it has also raised urgent questions about privacy, consent, and the misuse of one's digital image. One example of this is the rise of AI-powered nudification applications. Nudification applications enable users to artificially generate intimate imagery of someone without their consent, and without any technical skills or Photoshop expertise.[2] Advancements in AI have made it easier, faster, and cheaper (as low as $0.31 per image, depending on features and form of purchase) to produce this kind of harmful content.[3]

Legal scholar Mary Anne Franks poignantly articulated similar concerns in her 2011 article "Unwilling Avatars" about the misuse of images of women on internet:

> "What happens, then, when individuals are not in control of their online embodiment? What if one is embodied against one's will, in places that one never chooses to enter, in ways one never consented to be shown, in graphic and vicious detail for all the world to see and which may be impossible to erase?"[4]

This paper addresses these critical questions by focusing on the implications of AI-generated nonconsensual intimate imagery (NCII). It has two primary objectives: first, to "deblackbox" the technical workings of nudification systems, revealing how they operate through machine learning; and second, to provide a comprehensive overview of current research and policy developments. The purpose is to empower policymakers, technologists, and the public with a deeper understanding of these technologies and their broader societal impact.

What is NCII? Why is it important?

[2] Natalie Grace Brigham, Cassidy Gibson, Daniel Olszewski, Anna Crowder, Kevin R. B. Butler, Patrick Traynor, Elissa M. Redmiles, and Tadayoshi Kohno, *Analyzing the AI Nudification Application Ecosystem*, arXiv, November 14, 2024, https://doi.org/10.48550/arXiv.2411.09751.

[3] Ibid.

[4] Mary Anne Franks, "Unwilling Avatars: Idealism and Discrimination in Cyberspace," *Columbia Journal of Gender and Law* 20 (2011): 238.

According to research conducted in 2023, up to 98% of deepfake videos online contain sexually explicit intimate depictions.[5] Presumably, some of this content is consensual. However, this paper is focused on the ability to use AI tools as a form of abuse, and therefore is only interested in specifically nonconsensual content. This is not meant to be a stance on the ethics of pornography, but rather an examination of image-based sexual abuse (IBSA) facilitated by AI technologies.[6]

NCII represents a serious form of digital sexual abuse with severe consequences. Victim-survivors are predominantly women, and often experience real off-line harms as a consequence, including:

- Post-traumatic stress disorder
- Depression and anxiety
- Lowered self-esteem
- Social isolation
- Harm to reputation
- Potential loss of employment opportunities.[7]

Some have even been killed or taken their own lives as a direct result of this kind of abuse.[8]

This problem extends to minors. A recent study from the U.S. Department of Education found that 15% of public high school students–approximately 2.30 million

[5] Center for Democracy and Technology, *Civic Tech Fall Polling Research*, 2024, https://cdt.org/wp-content/uploads/2024/03/2024-03-21-CDT-Civic-Tech-Generative-AI-Survey-Research-final.pdf

[6] Travis L. Wagner and Ashley Blewer, "'The Word Real Is No Longer Real': Deepfakes, Gender, and the Challenges of AI-Altered Video," *Open Information Science* 3, no. 1 (2019): 32–46, https://doi.org/10.1515/opis-2019-0003.; Natalie Grace Brigham, Cassidy Gibson, Daniel Olszewski, Anna Crowder, Kevin R. B. Butler, Patrick Traynor, Elissa M. Redmiles, and Tadayoshi Kohno, *Analyzing the AI Nudification Application Ecosystem*.

[7] Natalie Grace Brigham, Cassidy Gibson, Daniel Olszewski, Anna Crowder, Kevin R. B. Butler, Patrick Traynor, Elissa M. Redmiles, and Tadayoshi Kohno, *Analyzing the AI Nudification Application Ecosystem*.

[8] Qiwei Li, Shihui Zhang, Andrew Timothy Kasper, Joshua Ashkinaze, Asia A. Eaton, Sarita Schoenebeck, and Eric Gilbert, *Reporting Non-Consensual Intimate Media: An Audit Study of Deepfakes*, arXiv, September 18, 2024, https://arxiv.org/abs/2409.12138.

public high school students–have experience with AI-generated NCII.[9] The involvement of minors complicates the issue, introducing additional legal and ethical dimensions:

1. The generated content constitutes CSAM (child sexual abuse material). CSAM is federally illegal, and scholars have argued that the same laws should apply to AI-generated CSAM.[10]
2. The victim-survivors are minors, and often don't receive adequate support from their schools or communities.[11]
3. The person generating this content is also a minor, which can complicate thinking about legal remedies, as will be discussed later in this paper.

Beyond individual impacts, the proliferation of nonconsensual imagery undermines digital trust, perpetuates gender-based violence, and poses significant legal and ethical challenges. Addressing NCII is crucial to creating a safer, more equitable online and offline environment.

Deblackboxing AI nudification systems

Through the use of computer vision and generative AI techniques, these tools can generate a naked image of someone (who is otherwise clothed) in a picture or video.[12] While an AI nudification system differs from a text-to-image generation system like DALL-E, they have many similarities in terms of their dependency on

[9] Center for Democracy and Technology, *Civic Tech Fall Polling Research*, 2024.

[10] David Thiel, Melissa Stroebel, and Rebecca Portnoff, *Generative ML and CSAM: Implications and Mitigations*, Stanford Internet Observatory and Thorn, June 24, 2023, https://stacks.stanford.edu/file/druid:jv206yg3793/20230624-sio-cg-csam-report.pdf.

[11] Center for Democracy and Technology, *Civic Tech Fall Polling Research*, 2024.

[12] Natalie Grace Brigham, Cassidy Gibson, Daniel Olszewski, Anna Crowder, Kevin R. B. Butler, Patrick Traynor, Elissa M. Redmiles, and Tadayoshi Kohno, *Analyzing the AI Nudification Application Ecosystem*.

training data, their reliance on pattern recognition, and their use of generative adversarial networks.

First, the AI system must recognize and interpret images as data, a process commonly referred to as "computer vision." "Computer vision" does not mean that a computer system can literally "see." Instead, it refers to a process that breaks down an image into computable data that the AI system can then analyze. This process starts with a digital image. Digital images are the result of light being transformed into numerical data by a camera's light-sensitive chip.[13] This data represents the colors and brightness of each pixel, creating a grid that forms a digital image.[14] By arranging these numbers in a grid, the camera creates a digital picture that software can analyze or change.[15]

These systems alter the numerical data encoded in digital images to simulate skin in place of clothing. This process relies on AI models trained on large datasets, enabling the system to make educated "guesses" that produce fake but realistic-looking pictures.[16] These predictions are informed by mathematical optimization and pattern recognition, allowing the AI to generate seamless manipulations by altering regions of the image while keeping the rest intact.[17] Pattern recognition plays a crucial role in this process. AI models, trained on vast datasets of images of clothed and unclothed bodies, learn to identify specific elements like clothing and skin. Using these patterns, the models generate predictive outputs that enable the software to replace clothing with simulated skin in a way that appears realistic and cohesive.

[13] Ron White, *How Digital Photography Works*, 2nd ed. (Indianapolis: Que Publishing, 2007).
[14] Ibid.
[15] Ibid.
[16] Ibid.
[17] Jay Peters, "AI Is Confusing — Here's Your Cheat Sheet," *The Verge*, July 22, 2024, https://www.theverge.com/24201441/ai-terminology-explained-humans.

Modern AI systems, such as those in nudification apps, rely on convolutional neural networks (CNNs) (or similar architectures), a type of artificial neural network specialized for image analysis. CNNs analyze images by focusing on smaller details like edges and textures to gradually build a complete understanding of the image.[18] More specifically, CNNs “break down an image” by applying filters, also called “kernels” to detect patterns like edges, shapes, and textures through progressively abstract representations.[19] These patterns are then used to generate plausible skin textures, creating manipulated images that appear realistic.

Additionally, recent advances have introduced hybrid architectures like Vision Transformers, which may play a role in these applications by enhancing image analysis and feature detection.[20] These systems integrate attention mechanisms into these systems, which focus on critical areas of an image, such as transitions between skin and clothing.[21] By prioritizing these regions, attention mechanisms improve the seamlessness and believability of manipulated images, effectively addressing challenges like blending textures and matching lighting conditions.[22]

Realistic visual content is further enabled by generative adversarial networks (GANs), which consist of two components: a generator and a discriminator. The generator creates new image data, and the discriminator evaluates the realism of the

[18] Andrej Karpathy, “What a Deep Neural Network Thinks About Your #Selfie,” *Karpathy Blog*, October 25, 2015, https://karpathy.github.io/2015/10/25/selfie/

[19] Xia Zhao, Limin Wang, Yufei Zhang, Xuming Han, Muhammet Deveci, and Milan Parmar, “A Review of Convolutional Neural Networks in Computer Vision,” *Artificial Intelligence Review* 57, no. 4 (2024): 99, https://doi.org/10.1007/s10462-024-10721-6.

[20]Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, “An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale,” *arXiv*, October 22, 2020, https://arxiv.org/abs/2010.11929.

[21] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao, “Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction Without Convolutions,” *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, 568–578, https://doi.org/10.1109/ICCV48922.2021.00061.

[22] Ibid.

generated output.[23] These two networks work together to model visual patterns, like skin tone and texture.[24] For example, the generator would synthesize details to add to the original image, while the discriminator would make sure the additions look natural. This process enables nudification apps to produce detailed and realistic outputs.[25]

Finally, the success of these systems depends heavily on training data. Training data provides the foundation for pattern recognition, enabling AI to differentiate features like clothing and skin. Diverse datasets, containing varied body types, poses, and lighting, help the system detect and replicate patterns. Without sufficient training data, the system's predictions lead to unrealistic outputs. CNNs and GANs rely on this data to refine their ability to identify features and recreate missing details.[26] The data these systems are trained on are fundamental to the app's functionality. In the case of nudification apps specifically, these AI systems are presumably trained on a range of content, including clothed and unclothed images of individuals, so that the model is able to predict what "nudified" content would look like.

To bring together the concepts discussed above, here is a step-by-step overview of how these nudification apps work:

1. Image input:
    a. User uploads a photo, which the app processes into digital data that can be analyzed.
2. Pattern recognition:

[23] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems* 27 (2014), https://arxiv.org/abs/1406.2661.

[24] Andrej Karpathy, "What a Deep Neural Network Thinks About Your #Selfie," *Karpathy Blog*.

[25] *The Data Scientist*, "The Ethical and Societal Implications of Deep Nude Technology," April 9, 2024, https://thedatascientist.com/the-ethical-and-societal-implications-of-deep-nude-technology/.

[26] Andrej Karpathy, "What a Deep Neural Network Thinks About Your #Selfie," *Karpathy Blog*.

    a. The AI system scans the image to detect and segment areas representing clothing versus skin. This step relies on machine learning models trained to identify patterns such as texture, color, and edges.

3. Predictive modeling:
    a. The app uses its trained model to predict what the body might look like beneath the clothing. This involves generating new data based on prior patterns and contextual cues from the image.
4. Image manipulation:
    a. The system (the generator network of a GAN) replaces the pixels corresponding to clothing with newly generated pixels that simulate realistic skin tone, texture, and shading, ensuring consistency with the surrounding areas.
5. Refinement and rendering:
    a. Advanced algorithms (the discriminator network of a GAN) smooth transitions between altered and unaltered areas, blending the changes into the original image to create a realistic and seamless final product.

The role of open-sourced AI models in the AI nudification app ecosystem

Another important technical element of the systems behind AI-generated NCII is open-source AI frameworks.[27] The concept of open-source software has a long and important history in the development of technology.[28] It involves the free and open sharing of the details of software projects to encourage collaboration, modifications,

[27] Sunny Gandhi and Adam Billen, “The US Senate’s Passage of the TAKE IT DOWN ACT Is Progress on an Urgent, Growing Problem,” *Tech Policy Press*, February 21, 2025, https://www.techpolicy.press/the-us-senates-passage-of-the-take-it-down-act-is-progress-on-an-urgent-growing-problem/.

[28] Elizabeth Seger and Bessie O'Dell, *Open Horizons: Exploring Nuanced Technical and Policy Approaches to Openness in AI* (London: Demos, September 2024), https://demos.co.uk/wp-content/uploads/2024/08/Mozilla-Report_2024.pdf.

and improvements.[29] The practice of open-sourcing has even helped strengthen security and guard against misuse, in addition to increasing collaboration between sectors and increasing transparency.[30] However, while open-source software has several benefits, many of the nudification websites that exist today have been built on top of open-source models.[31]

Open-sourcing allows anyone to modify the model, making it easier for bad actors to bypass safeguards and introduce dangerous capabilities without oversight.[32] Even in closed AI models, adversarial users have repeatedly found ways to bypass safeguards through methods like prompt injection and jailbreaking, which enables misuse.[33] In open-source models, where developers lack control over fine-tuning and deployment, these vulnerabilities become significantly harder to detect and prevent.[34] Even models with mechanisms in place to disable NSFW content can be removed or bypassed.

To understand what it means to be open-sourced, we should return to the basics of how a generative AI model responds to a user prompt. When an AI model processes a user prompt, and responds with an output, the contents of that output are

[29] Ibid.

[30] Elizabeth Seger, Noemi Dreksler, Richard Moulange, Emily Dardaman, Jonas Schuett, K. Wei, Christoph Winter, Mackenzie Arnold, Seán Ó hÉigeartaigh, Anton Korinek, Markus Anderljung, Ben Bucknall, Alan Chan, Eoghan Stafford, Leonie Koessler, Aviv Ovadya, Ben Garfinkel, Emma Bluemke, Michael Aird, Patrick Levermore, Julian Hazell, and Abhishek Gupta, *Open-Sourcing Highly Capable Foundation Models: An Evaluation of Risks, Benefits, and Alternative Methods for Pursuing Open-Source Objectives*, Centre for the Governance of AI, September 29, 2023, https://cdn.governance.ai/Open-Sourcing_Highly_Capable_Foundation_Models_2023_GovAI.pdf.

[31] Natalie Grace Brigham, Cassidy Gibson, Daniel Olszewski, Anna Crowder, Kevin R. B. Butler, Patrick Traynor, Elissa M. Redmiles, and Tadayoshi Kohno, *Analyzing the AI Nudification Application Ecosystem*.

[32] Markus Anderljung and Anton Korinek, "Frontier AI Regulation: Safeguards Amid Rapid Progress," *Centre for the Governance of AI*, January 4, 2024, https://www.governance.ai/research-paper/frontier-ai-regulation-safeguards-amid-rapid-progress.

[33] Ibid.

[34] Ibid.

determined by a series of numerical parameters that make up the model.[35] These are commonly referred to as the model's weights, which are determined by the model's training data.[36] These weights determine a model's behavior and therefore the responses that users get from the model.[37] If a model's weights are released to the public, they can be customized beyond the control of the developer.[38] The original model developer no longer has insight into what the model is being used for or how it has been modified.[39] Furthermore, developers cannot limit access to the weights or the manipulated model after the fact.

In a closed AI model, the developer or administrator retains control over the system and can monitor misuse and disable abusive accounts.[40] In an unsecured, open-source model, the activities of the end user cannot be controlled or limited by the developer, or anyone for that matter.[41] This also makes it very difficult to detect if malicious actors have removed safety features or other guards meant to ensure safe AI deployment.[42] Not only can these safeguards be removed, but a malicious actor could manipulate a model's weights to maximize for abusive and harmful activity, such as NCII generation.[43] More specifically, once these models are trained to generate intimate imagery, they can proliferate with minimal oversight (Gandhi & Billen, 2025). This means that shutting down one AI nudification app may not be enough (Gandhi &

[35] National Telecommunications and Information Administration, *Dual-Use Foundation Models with Widely Available Model Weights* (Washington, DC: U.S. Department of Commerce, July 2024), https://www.ntia.gov/sites/default/files/publications/ntia-ai-open-model-report.pdf.

[36] Ibid.

[37] Ibid.

[38] Ibid.

[39] Ibid.

[40] David Evan Harris, "How to Regulate Unsecured 'Open-Source' AI: No Exemptions," *Tech Policy Press*, December 4, 2023, https://techpolicy.press/how-to-regulate-unsecured-opensource-ai-no-exemptions/.

[41] Ibid.

[42] Ibid.

[43] Ibid.

Billen, 2025). As Gandhi and Billen say in their Tech Policy Press article on the topic: “countless other copies already exist and are circulating” (2025).

The socio-technical context of AI nudification applications

This harmful use of generative AI technologies is not happening in isolation, but rather is part of a larger context of gender-based violence: “to the extent that they are used without the consent of the person in the image, their use is embedded in the sociocultural context of sexual abuse.”[44] A walkthrough study of the user interface of 20 websites with AI-based nudification tools found that 19 out of 20 websites explicitly specialize in the “nudification” of “women” and “girls,” which aligns with previous research done on non-AI generated IBSA, which found that over 90% of the content on “revenge porn” websites was of women.[45] The specific (and mostly exclusive) mention of “women” and “girls” demonstrates how these tools embody broader cultural ideas about the objectification and exploitation of women and girls. It is not the tools themselves that are inherently harmful, but rather the socio-political systems in which generative AI systems are created that encourages their use in this way.

AI-generated nudification content further perpetuates the dynamics of IBSA and gender-based violence already present in our culture. This study also found that some of these sites allow users to generate “nudification” content *without* needing to confirm that they are 18 years old or older.[46] Additionally, only some sites mention the subject of the image needing to be 18, and for those that do mention this, it’s only mentioned

[44] Natalie Grace Brigham, Cassidy Gibson, Daniel Olszewski, Anna Crowder, Kevin R. B. Butler, Patrick Traynor, Elissa M. Redmiles, and Tadayoshi Kohno, *Analyzing the AI Nudification Application Ecosystem*. P.3.

[45] Ibid: 7.

[46] Ibid.

in the site's terms of service, as opposed to being prominently displayed.[47] Furthermore, they found that only *half* of these websites say that users must obtain “consent” of the subject of the image before “undressing” them.[48] Even fewer of these sites ask for the user to confirm that they have acquired said consent.[49] These technologies were designed with little to no regard for the safety of those depicted in the images, or, arguably, for the safety of those generating the images, especially when the person generating the image is a minor.

The legal & political context of AI nudification apps

Efforts to regulate AI generated NCII have spread globally. Legislative approaches to address this issue vary across jurisdictions, with initiatives at federal, state, and international levels. One notable example is in San Francisco, where City Attorney David Chiu has filed a lawsuit against the 16 most popular websites that provide an AI-generated “nudification” service for a small fee.[50] The 16 websites that are targeted by the lawsuit have been visited 200 million times in the first half of 2024, according to Mr. Chiu.[51] The lawsuit alleges that the businesses behind these websites “broke numerous state laws against fraudulent business practices, nonconsensual pornography and the sexual abuse of children.”[52]

This lawsuit differs from other attempts to criminalize AI-generated & nonconsensual sexually explicit depictions of minors because it asks specifically for a

[47] Ibid.
[48] Ibid.
[49] Ibid.
[50] Heather Knight, “San Francisco Moves to Lead Fight Against Deepfake Nudes,” *The New York Times*, August 15, 2024, https://www.nytimes.com/2024/08/15/us/deepfake-pornography-lawsuit-san-francisco.html.
[51] Ibid.
[52] Associated Press, “San Francisco Files First-of-Its-Kind Lawsuit to Tackle AI Deepfake Nudes,” *Politico*, August 17, 2024, https://www.politico.com/news/2024/08/17/san-francisco-lawsuit-ai-deepfake-nudes-00174487.

judge to shut down these types of sites all together, while other lawsuits seek to go after the individuals creating and distributing the images.[53] Mr. Chiu argues that going after individuals, as opposed to shutting down the tools, is less effective, since "once the images are circulating, it is nearly impossible to determine [who] created them, making it very difficult for the women to successfully sue."[54]

At the federal level, there are two significant proposed laws that seek to address and remedy this kind of harm: The TAKE IT DOWN Act (Tools to Address Known Exploitation by Immobilizing Technological Deepfakes on Websites and Networks Act)[55] and the DEFIANCE Act (Digital Exploitation Free Interventions and Accountability for Nonconsensual Exploitation).[56] The TAKE IT DOWN Act, supported by technology advocacy groups, seeks to establish a national framework mandating the removal of explicit nonconsensual imagery by social media platforms upon request.[57] This is particularly important as there is currently no particularly effective method for victim-survivors to advocate for the removal of the content.[58] Since the start of 2025, Senators Ted Cruz and Amy Klobuchar have reintroduced the TAKE IT DOWN Act. As of April 8, 2025, it cleared the House Energy and Commerce

[53]Heather Knight, "San Francisco Moves to Lead Fight Against Deepfake Nudes," *The New York Times*.

[54] Ibid.

[55] U.S. Congress, Senate, *TAKE IT DOWN Act*, S.4569, 118th Cong., 2nd sess., introduced in Senate June 18, 2024, https://www.congress.gov/bill/118th-congress/senate-bill/4569.

[56] U.S. Congress, Senate, *Disrupt Explicit Forged Images and Non-Consensual Edits Act of 2024*, S.3696, 118th Cong., 2nd sess., introduced January 30, 2024, https://www.congress.gov/bill/118th-congress/senate-bill/3696.

[57] Maria Curi, "Exclusive: Tech Groups Back Bills to Fight Harmful AI Imagery," *Axios*, December 5, 2024, https://www.axios.com/pro/tech-policy/2024/12/05/tech-groups-back-bills-to-fight-harmful-ai-imagery

[58] Qiwei Li, Shihui Zhang, Andrew Timothy Kasper, Joshua Ashkinaze, Asia A. Eaton, Sarita Schoenebeck, and Eric Gilbert, *Reporting Non-Consensual Intimate Media: An Audit Study of Deepfakes*.

Committee and has unanimously passed the Senate.[59] As of April 29th, 2025, the TAKE IT DOWN Act has passed.[60]

Meanwhile, the DEFIANCE Act, passed by the Senate in July 2024, seeks to create enforceable protections against this kind of content, requiring companies to implement safeguards to prevent its creation and dissemination. The Act also mandates a federal grant program to support victims of digital exploitation.[61]

Several U.S. states have introduced or enacted laws targeting AI-generated deepfake content. These laws often build upon existing frameworks for addressing revenge pornography or other nonconsensual imagery. For example, recent legislation in Texas and California specifically criminalizes the creation and distribution of sexually explicit deepfakes, with additional penalties for content targeting minors.[62] Georgia has established a committee to explore the societal and legal ramifications of deepfakes, signaling a broader trend of state-level policy experimentation.[63]

Outside the United States, the European Union offers a contrasting regulatory model. The EU's Artificial Intelligence Act incorporates provisions targeting harmful AI applications, including nonconsensual deepfake imagery, through a risk-based

[59] U.S. Senate Committee on Commerce, Science, & Transportation, "Sens. Cruz and Klobuchar's Bipartisan TAKE IT DOWN Act Clears House Committee," April 8, 2025, https://www.commerce.senate.gov/2025/4/sens-cruz-and-klobuchar-s-bipartisan-take-it-down-act-clears-house-committee.

[60] Barbara Ortutay, "Take It Down Act, Addressing Nonconsensual Deepfakes and 'Revenge Porn,' Passes. What Is It?" *AP News*, April 29, 2025, https://apnews.com/article/take-it-down-deepfake-trump-melania-first-amendment-741a6e525e81e5e3d8843aac20de8615

[61] Alexandria Ocasio-Cortez, "Ocasio-Cortez Statement on Senate Passage of the DEFIANCE Act," *Office of Congresswoman Alexandria Ocasio-Cortez*, July 23, 2024, https://ocasio-cortez.house.gov/media/press-releases/ocasio-cortez-statement-senate-passage-defiance-act

[62] Safia Samee Ali, "What State Laws Protect Kids Against AI-Generated Deepfakes?" *NewsNation*, November 26, 2024, https://www.newsnationnow.com/business/tech/laws-kids-ai-generated-deepfakes/.

[63] Thomas Wheatley, "Georgia Lawmakers Want to Hold Deepfake AI Developers Accountable," *Axios*, December 4, 2024, https://www.axios.com/local/atlanta/2024/12/04/georgia-ai-study-committee-deepfakes.

framework.[64] This legislation emphasizes transparency, requiring providers of AI tools to implement disclosure measures and to ensure their systems cannot be easily exploited for malicious purposes.[65] The EU's approach reflects a broader emphasis on proactive governance and systemic safeguards.

Challenges to policy addressing AI-generated nudification content

Despite these advancements, challenges persist in regulating AI-generated nudification content. The rapid development of deepfake technology consistently outpaces legislative processes, creating significant enforcement gaps. The decentralized nature of the internet further complicates accountability, making it extraordinarily difficult for victim-survivors to identify the origin of generated content, track the spread of manipulated images, and pursue meaningful legal recourse.

State-level approaches to regulating AI-generated nonconsensual intimate imagery (NCII) have created a patchwork of inconsistent protections.[66] Jurisdictions vary widely in their definitions of nonconsensual sexually explicit deepfakes, the levels of intent required to qualify as a violation, exemptions, punishment levels, and available victim remedies. This fragmentation means individuals are protected differently depending on their geographic location, creating an inherently inequitable system of digital safety.[67]

I. Challenges to holding perpetrators or platforms accountable

[64] Mauro Fragale and Valentina Grilli, "Deepfake, Deep Trouble: The European AI Act and the Fight Against AI-Generated Misinformation," *Columbia Journal of European Law*, November 11, 2024, https://cjel.law.columbia.edu/preliminary-reference/2024/deepfake-deep-trouble-the-european-ai-act-and-the-fight-against-ai-generated-misinformation/.

[65] Felipe Romero-Moreno, "Generative AI and Deepfakes: A Human Rights Approach to Tackling Harmful Content," *International Review of Law, Computers & Technology* 38, no. 3 (2024): 297–326, https://doi.org/10.1080/13600869.2024.2324540.

[66] Meeka Bondy, John Delaney, and Jeff Ong, "AI-Generated Deepfakes and the Emerging Legal Landscape," *Age of Disruption* (blog), Perkins Coie LLP, April 15, 2024, https://perkinscoie.com/insights/blog/ai-generated-deepfakes-and-emerging-legal-landscape.

[67] Ibid.

Existing legal avenues prove insufficient for addressing AI-generated NCII. For example, defamation could be used as a response to AI-generated NCII, but some legal experts argue that these laws are insufficient to address these harms. This is because defamation is considered a civil offense, rather than a criminal offense, so it provides only financial remedies that cannot effectively remove content from platforms or prevent future content creation.[68] These limitations mean that victim-survivors are left with minimal recourse, unable to address the immediate harm or long-term reputational damage caused by such imagery. For these reasons, there is significant demand for federal legislation to address these gaps.

The 1996 Communications Decency Act (Section 230) presents another critical challenge by providing online intermediaries immunity from legal liability for user-generated content. Section 230 states that online intermediaries, like social media platforms, are not legally considered to be the publisher or speaker of third-party content.[69] This legislation means that social media platforms face no consequences for hosting harmful content, leaving victim-survivors unable to hold these platforms accountable. The anonymity of perpetrators further compounds this issue, making legal pursuit nearly impossible under the current framework.

II. Challenges to balancing regulation with Free Speech

The challenge of regulating AI-generated nonconsensual intimate imagery is further complicated by the delicate balance between protecting individual dignity and preserving free speech. This is especially difficult because the question of whether AI-generated outputs are constitutionally protected speech (and, if so, whether AI-generated NCII would enjoy the same privileges) is unanswered. Legal scholar Mary

[68] Caroline Quirk, "The High Stakes of Deepfakes: The Growing Necessity of Federal Legislation to Regulate This Rapidly Evolving Technology," *Princeton Legal Journal*, June 20, 2023, https://legaljournal.princeton.edu/the-high-stakes-of-deepfakes-the-growing-necessity-of-federal-legislation-to-regulate-this-rapidly-evolving-technology/.

[69] James Grimmelmann, *Internet Law: Cases and Problems*, 14th ed. (Semaphore Press, 2024).

Anne Franks (2011) argues that unchecked "liberty" to harass or exploit others online undermines the broader liberty of marginalized groups, particularly women, to participate in digital spaces. She contends that "regulation and reform are necessary to balance the equation and create a cyberspace that maximizes liberty for all groups."[70] A failure to regulate such content perpetuates systemic inequalities, as women disproportionately bear the brunt of online sexual abuse. However, this requires carefully crafted legislation that protects individuals from harmful content without enabling inappropriate censorship.[71]

The American Civil Liberties Union (ACLU), for instance, has opposed legislation that broadly targets deepfakes, viewing such laws as potentially infringing on free expression.[72] The recent judicial overturn of California's election-related deepfake law further illustrates the complex constitutional challenges. The law in question allowed individuals to sue over AI-generated false or deceptive content related to elections.[73] A federal judge overturned the law due to concerns that it likely infringes on people's First Amendment rights, which protects speech like political satire.[74] While this specific example differs from non-consensual intimate imagery, it demonstrates the judicial system's vigilance in protecting First Amendment rights. Consequently, any effective legislative approach must thread a narrow path: protecting vulnerable individuals from digital abuse while simultaneously preserving fundamental free speech protections.

III. Challenges to penalizing minors

[70] Mary Anne Franks, "Unwilling Avatars: Idealism and Discrimination in Cyberspace," *Columbia Journal of Gender and Law* 20 (2011): 249.
[71] Arthur Holland Michel, "The ACLU Fights for Your Constitutional Right to Make Deepfakes," *Wired*, July 24, 2024, https://www.wired.com/story/aclu-artificial-intelligence-deepfakes-free-speech/.
[72] Ibid.
[73] Associated Press, "Judge Blocks New California Law Cracking Down on Election Deepfakes," *AP News*, October 2, 2024, https://apnews.com/article/california-deepfake-election-law-ee5a3d7cba3e9f5caddf91b127e4938a.
[74] Ibid.

Another significant difficulty with legislating this issue is that the perpetrators are sometimes minors themselves.[75] Some argue that, while law enforcement is an important piece of addressing this issue, it isn't always the right answer, as imposing felonies on minors can threaten to reinforce the school-to-prison pipeline.[76] Furthermore, punitive measures fail to address underlying factors contributing to this behavior, such as a lack of digital literacy education or a lack of comprehensive sexual education that promotes healthy attitudes towards consent and respect.

Potential solutions

When thinking about solutions to the problem of AI-generated nonconsensual intimate material, there are two main problems to address: 1. What can we do to stop (or at least sharply decrease) the availability of tools that facilitate the generation of this kind of content? And 2. How do we deal with this kind of content that already exists, and is potentially being circulated?

I. Illegal content in training data

In terms of the specific problem of these AI tools being used to generate CSAM, one potential technical solution is to remove known CSAM from training data and models. A report from 2023 from the Stanford Internet Observatory and Thorn found that rapid developments in generative Ai made it possible to create images that facilitate child sexual exploitation using open source AI image generation models.[77] A follow-up report found that one specific image generating AI tool, Stable Diffusion,

[75] Kaylee Williams, "Minors Are on the Frontlines of the Sexual Deepfake Epidemic—Here's Why That's a Problem," *Tech Policy Press*, October 10, 2024, https://www.techpolicy.press/minors-are-on-the-frontlines-of-the-sexual-deepfake-epidemic-heres-why-thats-a-problem/.

[76] Ibid.

[77] David Thiel, Melissa Stroebel, and Rebecca Portnoff, *Generative ML and CSAM: Implications and Mitigations*, Stanford Internet Observatory and Thorn.

was trained on a dataset that included CSAM.[78] The specific dataset in question is LAION-5B, which researchers found to contain "thousands of illegal images" as recently as late 2023.[79] According to the report: "while the amount of CSAM present does not necessarily indicate that the presence of CSAM drastically influences the output of the model above and beyond the model's ability to combine the concepts of sexual activity and children, it likely does still exert influence," meaning that removing that data is key to creating a safer (even if not foolproof) system.[80]

Thiel recommends steps that can be taken to mitigate this problem and to prevent these incidents in the future, including removing this content from the models trained on this dataset.

The presence of CSAM in this dataset, which was "fed by essentially unguided crawling," demonstrates how, without intervention and guidance, AI systems and their capabilities are based on any data they are trained on–legal or not.[81] When that data includes everything on the internet, the system will act accordingly. This is why "cleaning up" training data is such an important step in developing AI systems. In terms of removing this material from the model, Thiel says this is a difficult task.[82] He says that the image and text embeddings for the images that match known CSAM could be removed, but that "it is unknown whether this would meaningfully affect the ability of the model to produce CSAM or to replicate the appearance of specific victims."[83]

While it's not clear whether removing the images from the model would solve the problem or not, Thiel recommends using caution when training models on images of children moving forward. He had previously proposed that "models trained on erotic

[78] David Thiel, *Identifying and Eliminating CSAM in Generative ML Training Data and Models* (Stanford, CA: Stanford Internet Observatory, December 2023), https://doi.org/10.25740/kh752sm9123.
[79] Ibid: 10.
[80] Ibid.
[81] Ibid: 2.
[82] Ibid.
[83] Ibid.

content not be trained on material depicting children," since this will limit "the ability of the model to conflate the two types of material."[84] He goes one step further, pointing out that "given the regulatory scrutiny regarding gathering data on children (e.g. COPPA), images of children should arguably be excluded from generalized training sets entirely."[85] COPPA refers to The Children's Online Privacy Protection Act, a U.S. federal law designed to protect the privacy of children under the age of 13 by requiring online services to obtain parental consent before collecting personal information from children.[86] It establishes strict requirements for data collection, storage, and use, in an attempt to ensure that online platforms prioritize child safety and privacy.[87]

II. Potential detecting technologies

Another potential technical solution is using detection techniques to detect deepfakes and CSAM. Some AI systems can be trained to recognize inconsistencies in visual data (like unnatural lighting, distorted pixel pattern, or facial motion inconsistencies) that are typical of AI-generated content.[88] However, the rapid advancement of AI-generated deepfake technology can outpace detection methods. Presumably, as generative AI systems continue improving, its imperfections will become increasingly difficult to detect.

However, there are more reliable ways to detect known CSAM, which could possibly be expanded to include AI-generated CSAM & previously identified NCII to

[84] Ibid: 11.

[85] Ibid.

[86] Federal Trade Commission, "Children's Online Privacy Protection Rule ('COPPA')," accessed May 6, 2025, https://www.ftc.gov/legal-library/browse/rules/childrens-online-privacy-protection-rule-coppa.

[87] Ibid.

[88]Nguyen, Thanh Thi, Quoc Viet Hung Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Thien Huynh-The, Saeid Nahavandi, Thanh Tam Nguyen, Quoc-Viet Pham, and Cuong M. Nguyen. "Deep Learning for Deepfakes Creation and Detection: A Survey." *Computer Vision and Image Understanding* 223 (2022): 103525. https://doi.org/10.1016/j.cviu.2022.103525; Agarwal, Shruti, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. "Protecting World Leaders Against Deep Fakes." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (CVPRW), 2019. https://farid.berkeley.edu/downloads/publications/cvpr19/cvpr19a.pdf.vgl.ict.usc.edu+2.

stop its spread on social media platforms. One important tool is Microsoft PhotoDNA, which uses hash-matching technology to convert images into unique hashes.[89] Hashes work essentially as a digital fingerprint, which can then be compared against a database of known CSAM hashes to identify and remove illegal content.[90] By using this hashing method, PhotoDNA enables the moderation of CSAM without requiring a human to actually view the CSAM in question By including known AI-generated CSAM and NCII into PhotoDNA, it could make it easier for platforms to remove the content. The only problem with this type of approach is that it works well for images that are already known to be harmful or illegal, but would not be helpful in detecting new, unknown content.

III. The role of platforms

Another system that helps facilitate the removal of NCII on platforms and could be used for AI-generated NCII is StopNCII.org, which provides tools for individuals to request the removal of NCII across platforms.[91] Organizations like StopNCII collaborate with social media platforms to ensure quicker response times.[92] Now that it has passed, the TAKE IT DOWN Act will enshrine this kind of process into law, and mandate that social media platforms respond to a victim-survivor's request for the removal of AI-generated NCII in a timely manner.[93] By fostering collaboration between platforms, policymakers, and advocacy groups, the TAKE IT DOWN Act ideally will encourage a more unified and effective approach to content moderation, so that these images don't exist in cyberspace permanently.

IV. The role of culture

[89] Microsoft, "PhotoDNA," accessed May 6, 2025, https://www.microsoft.com/en-us/photodna.
[90] Ibid.
[91] StopNCII.org, "About Us," accessed May 6, 2025, https://stopncii.org/about-us/.
[92] Ibid.
[93] U.S. Congress, Senate, *TAKE IT DOWN Act*, S.4569, 118th Cong., 2nd sess.

In order to address the problem of IBSA (image bases sexual abuse) at a more fundamental level, beyond any technological tool, there needs to be a cultural change around both digital literacy and sex education. Comprehensive sexuality education (CSE) programs, that include education about digital literacy, can promote respect, consent, and accountability, while preventing behaviors that lead to NCII and other forms of digital sexual harassment and abuse. CSE programs can challenge and reshape cultural norms like patriarchal attitudes and toxic masculinity that contribute to gender- based violence.[94] Addressing these attitudes and the problem of AI-generated NCII/CSAM specifically could create a cultural shift toward more respectful gender dynamics and less sexual harm. A first step could include talking about preventing AI-generated NCII in schools, so that students understand the harms and consequences of this sort of behavior. This is an area of focus that warrants further research and exploration.

Conclusion

The issue of AI-generated NCII represents how, while technological innovations can move society forward, they can also enable some of society's darkest behavior. This paper has highlighted the technical mechanisms that enable these systems, the legal gaps that complicate accountability, and the cultural factors that perpetuate harm. Addressing these challenges requires a multi-pronged approach: improving technical safeguards, advancing targeted legislation, and fostering cultural change through digital literacy and comprehensive sex education. Looking ahead, reinforcement learning could be a promising avenue for mitigating the generation of AI-generated NCII.[95] By incorporating reinforcement learning algorithms that penalize

[94] Michael Flood and Bob Pease, "Factors Influencing Attitudes to Violence Against Women," *Trauma, Violence & Abuse* 10, no. 2 (2009): 125–42, https://doi.org/10.1177/1524838009334131.

[95] Leo Gao, John Schulman, and Jacob Hilton, "Scaling Laws for Reward Model Overoptimization," *arXiv*, October 19, 2022, https://doi.org/10.48550/arXiv.2210.10760.

harmful outputs and reward ethical behavior in AI systems, developers could create models that are better aligned with societal values and legal standards.[96] Exploring reinforcement learning as a preventative mechanism across systems warrants further research, as it could play a crucial role in reducing the production of nonconsensual content and ensuring AI tools are used responsibly. This is something that large, multiuse generative AI companies are doing, like OpenAI, but there might be a way for the AI community to agree on a standard of reinforcement learning across the board to prevent any tools from being able to be used in this way. In addition to exploring technical improvements through reinforcement learning, some developers have implemented other measures intended to prevent misuse. For example, Google employs a combination of automated detection tools and trained reviewers to deter, detect, remove, and report CSAM, including AI-generated CSAM, in addition to collaborating with organizations like the Internet Watch Foundation and Thorn to address these risks.[97]

Moving forward, interdisciplinary collaboration between technologists, policymakers, and educators is crucial to develop solutions that balance innovation with safety and respect. Policymakers must align efforts to create a federal legal framework, technologists must prioritize transparency and ethical design, and educators must equip individuals with the tools to navigate a world of technological innovation. The rapid evolution of AI demands that we protect against abuses that threaten autonomy and equity in order to ensure a safer and more just future for all.

[96] Ibid.

[97]Google. *Progress Update: Responsible AI and Child Sexual Abuse and Exploitation Online*. 2025. https://static.googleusercontent.com/media/publicpolicy.google/en//resources/ai_responsibility_and_csae_en.pdf.

References:

Agarwal, Shruti, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. "Protecting World Leaders Against Deep Fakes." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (CVPRW), 2019. https://farid.berkeley.edu/downloads/publications/cvpr19/cvpr19a.pdf.

Ali, Safia Samee. "What State Laws Protect Kids Against AI-Generated Deepfakes?" *NewsNation*, November 26, 2024. https://www.newsnationnow.com/business/tech/laws-kids-ai-generated-deepfakes/.

Anderljung, Markus, and Anton Korinek. "Frontier AI Regulation: Safeguards amid Rapid Progress." January 4, 2024. https://www.governance.ai/research-paper/frontier-ai-regulation-safeguards-amid-rapid-progress.

Associated Press. "San Francisco Files First-of-Its-Kind Lawsuit to Tackle AI Deepfake Nudes." *Politico*, August 17, 2024. https://www.politico.com/news/2024/08/17/san-francisco-lawsuit-ai-deepfake-nudes-00174487

Associated Press, "Judge Blocks New California Law Cracking Down on Election Deepfakes," *AP News*, October 2, 2024, https://apnews.com/article/california-deepfake-election-law-ee5a3d7cba3e9f5caddf91b127e4938a.

Bondy, Meeka, John Delaney, and Jeff Ong. "AI-Generated Deepfakes and the Emerging Legal Landscape." *Age of Disruption* (blog). Perkins Coie LLP. April 15, 2024. https://perkinscoie.com/insights/blog/ai-generated-deepfakes-and-emerging-legal-landscape.

Brigham, Natalie Grace, Miranda Wei, Tadayoshi Kohno, and Elissa M. Redmiles. "Violation of My Body: Perceptions of AI-Generated Non-Consensual (Intimate) Imagery." *arXiv*, June 8, 2024. https://doi.org/10.48550/arXiv.2406.05520

Center for Democracy and Technology. *Civic Tech Fall Polling Research*. Washington, D.C.: Center for Democracy and Technology, 2024. https://cdt.org/wp-content/uploads/2024/03/2024-03-21-CDT-Civic-Tech-Generative-AI-Survey-Research-final.pdf.

Curi, Maria. "Exclusive: Tech Groups Back Bills to Fight Harmful AI Imagery." *Axios*, December 5, 2024. https://www.axios.com/pro/tech-policy/2024/12/05/tech-groups-back-bills-to-fight-harmful-ai-imagery.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. “An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale.” *arXiv*, October 22, 2020. https://arxiv.org/abs/2010.11929.

Dwyer, Maddy, and Elizabeth Laird. *Up in the Air: Educators Juggling the Potential of Generative AI with Detection, Discipline, and Distrust*. Washington, D.C.: Center for Democracy and Technology, March 2024. https://cdt.org/wp-content/uploads/2024/03/2024-03-21-CDT-Civic-Tech-Generative-AI-Survey-Research-final.pdf.

Federal Trade Commission. “Children’s Online Privacy Protection Rule (‘COPPA’).” Accessed May 6, 2025. https://www.ftc.gov/legal-library/browse/rules/childrens-online-privacy-protection-rule-coppa.

Finer, Lauren. “Senate Passes DEFIANCE Act Targeting Non-Consensual Intimate AI Deepfakes.” *The Verge*, July 24, 2024. https://www.theverge.com/2024/7/24/24205275/senate-passes-defiance-act-non-consensual-intimate-ai-deepfakes.

Flood, Michael, and Bob Pease. “Factors Influencing Attitudes to Violence Against Women.” *Trauma, Violence & Abuse* 10, no. 2 (2009): 125–142. https://doi.org/10.1177/1524838009334131.

Fragale, Mauro, and Valentina Grilli. “Deepfake, Deep Trouble: The European AI Act and the Fight Against AI-Generated Misinformation.” *Columbia Journal of European Law*, November 11, 2024. https://cjel.law.columbia.edu/preliminary-reference/2024/deepfake-deep-trouble-the-european-ai-act-and-the-fight-against-ai-generated-misinformation/.

Franks, Mary Anne. “Unwilling Avatars: Idealism and Discrimination in Cyberspace.” *Columbia Journal of Gender and Law* 20, no. 2 (2011): 224–261. https://repository.law.miami.edu/fac_articles/306.

Gao, Leo, John Schulman, and Jacob Hilton. “Scaling Laws for Reward Model Overoptimization.” *arXiv*, October 19, 2022. https://doi.org/10.48550/arXiv.2210.10760

Gandhi, Sunny, and Adam Billen. “The US Senate’s Passage of the TAKE IT DOWN ACT Is Progress on an Urgent, Growing Problem.” *Tech Policy Press*, February 21, 2025. https://www.techpolicy.press/the-us-senates-passage-of-the-take-it-down-act-is-progress-on-an-urgent-growing-problem/.

Gibson, Cassidy, Daniel Olszewski, Natalie Grace Brigham, Anna Crowder, Kevin R. B. Butler, Patrick Traynor, Elissa M. Redmiles, and Tadayoshi Kohno. “Analyzing the

AI Nudification Application Ecosystem." *arXiv*, November 14, 2024. https://doi.org/10.48550/arXiv.2411.09751

Graham, Michelle M. "Deepfakes: Federal and State Regulation Aims to Curb a Growing Threat." *Thomson Reuters*, June 26, 2024. https://www.thomsonreuters.com/en-us/posts/government/deepfakes-federal-state-regulation/

Grimmelmann, James. *Internet Law: Cases and Problems*. 14th ed. Semaphore Press, 2024.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative Adversarial Nets." In *Advances in Neural Information Processing Systems* 27, 2014. https://arxiv.org/abs/1406.2661.

Google. *Progress Update: Responsible AI and Child Sexual Abuse and Exploitation Online*. 2025. https://static.googleusercontent.com/media/publicpolicy.google/en//resources/ai_responsibility_and_csae_en.pdf.

Harris, David Evan. "How to Regulate Unsecured 'Open-Source' AI: No Exemptions." *Tech Policy Press*, December 4, 2023. https://techpolicy.press/how-to-regulate-unsecured-opensource-ai-no-exemptions/.

Michel, Arthur Holland. "The ACLU Fights for Your Constitutional Right to Make Deepfakes." *Wired*, July 24, 2024. https://www.wired.com/story/aclu-artificial-intelligence-deepfakes-free-speech/.

Karpathy, Andrej. "What a Deep Neural Network Thinks About Your #Selfie." *Karpathy Blog*. October 25, 2015. https://karpathy.github.io/2015/10/25/selfie/.

Knight, Heather. "San Francisco Moves to Lead Fight Against Deepfake Nudes." *The New York Times*, August 15, 2024. https://www.nytimes.com/2024/08/15/us/deepfake-pornography-lawsuit-san-francisco.html.

Microsoft. "PhotoDNA." Accessed May 6, 2025. https://www.microsoft.com/en-us/photodna.

National Telecommunications and Information Administration. *Dual-Use Foundation Models with Widely Available Model Weights*. Washington, DC: U.S. Department of Commerce, July 2024. https://www.ntia.gov/sites/default/files/publications/ntia-ai-open-model-report.pdf.

Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Nguyen, D. T., Huynh-The, T., Nahavandi, S., Nguyen, T. T., Pham, Q.-V., & Nguyen, C. M. (2022). Deep learning for deepfakes creation and detection: A survey. Computer Vision and Image Understanding, 223, 103525. https://doi.org/10.1016/j.cviu.2022.103525

Ocasio-Cortez, Alexandria. “Ocasio-Cortez Statement on Senate Passage of the DEFIANCE Act.” Office of Congresswoman Alexandria Ocasio-Cortez. July 23, 2024. https://ocasio-cortez.house.gov/media/press-releases/ocasio-cortez-statement-senate-passage-defiance-act

Ortutay, Barbara. “Take It Down Act, Addressing Nonconsensual Deepfakes and 'Revenge Porn,' Passes. What Is It?” *AP News*, April 29, 2025. https://apnews.com/article/741a6e525e81e5e3d8843aac20de8615

Peters, Jay. “AI Is Confusing — Here’s Your Cheat Sheet.” *The Verge*, July 22, 2024. https://www.theverge.com/24201441/ai-terminology-explained-humans.

Quirk, Caroline. “The High Stakes of Deepfakes: The Growing Necessity of Federal Legislation to Regulate This Rapidly Evolving Technology.” *Princeton Legal Journal*, June 20, 2023. https://legaljournal.princeton.edu/the-high-stakes-of-deepfakes-the-growing-necessity-of-federal-legislation-to-regulate-this-rapidly-evolving-technology/.

Romero-Moreno, Felipe. “Generative AI and Deepfakes: A Human Rights Approach to Tackling Harmful Content.” *International Review of Law, Computers & Technology* 38, no. 3 (2024): 297–326. https://doi.org/10.1080/13600869.2024.2324540.

Seger, Elizabeth, and Bessie O'Dell. *Open Horizons: Exploring Nuanced Technical and Policy Approaches to Openness in AI*. London: Demos, September 2024. https://demos.co.uk/wp-content/uploads/2024/08/Mozilla-Report_2024.pdf.

Seger, Elizabeth, Noemi Dreksler, Richard Moulange, Emily Dardaman, Jonas Schuett, K. Wei, Christoph Winter, Mackenzie Arnold, Seán Ó hÉigeartaigh, Anton Korinek, Markus Anderljung, Ben Bucknall, Alan Chan, Eoghan Stafford, Leonie Koessler, Aviv Ovadya, Ben Garfinkel, Emma Bluemke, Michael Aird, Patrick Levermore, Julian Hazell, and Abhishek Gupta. *Open-Sourcing Highly Capable Foundation Models: An Evaluation of Risks, Benefits, and Alternative Methods for Pursuing Open-Source Objectives*. Centre for the Governance of AI, September 29, 2023. https://cdn.governance.ai/Open-Sourcing_Highly_Capable_Foundation_Models_2023_GovAI.pdf.

StopNCII.org. “About Us.” Accessed May 6, 2025. https://stopncii.org/about-us/.

The Data Scientist. “The Ethical and Societal Implications of Deep Nude Technology.” April 9, 2024. https://thedatascientist.com/the-ethical-and-societal-implications-of-deep-nude-technology/.

Thiel, David. *Identifying and Eliminating CSAM in Generative ML Training Data and Models*. Stanford, CA: Stanford Internet Observatory, December 2023. https://doi.org/10.25740/kh752sm9123.

Thiel, David, Stroebel, Stroebel, and Portnoff, Rebecca. *Generative ML and CSAM: Implications and Mitigations*, Stanford Internet Observatory and Thorn, June 24, 2023, https://stacks.stanford.edu/file/druid:jv206yg3793/20230624-sio-cg-csam-report.pdf.

U.S. Senate Committee on Commerce, Science, & Transportation. "Sens. Cruz and Klobuchar's Bipartisan TAKE IT DOWN Act Clears House Committee." April 8, 2025. https://www.commerce.senate.gov/2025/4/sens-cruz-and-klobuchar-s-bipartisan-take-it-down-act-clears-house-committee.

U.S. Congress. Senate. *TAKE IT DOWN Act*. S.4569. 118th Cong., 2nd sess. Introduced in Senate June 18, 2024. https://www.congress.gov/bill/118th-congress/senate-bill/4569.

U.S. Congress, Senate, *Disrupt Explicit Forged Images and Non-Consensual Edits Act of 2024*, S.3696, 118th Cong., 2nd sess., introduced January 30, 2024, https://www.congress.gov/bill/118th-congress/senate-bill/3696.

Wagner, Travis L., and Ashley Blewer. "The Word Real Is No Longer Real: Deepfakes, Gender, and the Challenges of AI-Altered Video." *Open Information Science* 3, no. 1 (2019): 32–46. https://doi.org/10.1515/opis-2019-0003.

Wang, Wenhai, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. "Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction Without Convolutions." *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, 568–578. https://doi.org/10.1109/ICCV48922.2021.00061.

Wheatley, Thomas. "Georgia Lawmakers Want to Hold Deepfake AI Developers Accountable." *Axios*, December 4, 2024. https://www.axios.com/local/atlanta/2024/12/04/georgia-ai-study-committee-deepfakes.

White, Ron. *How Digital Photography Works*. 2nd ed. Indianapolis: Que Publishing, 2007.

Williams, Kaylee. "Minors Are on the Frontlines of the Sexual Deepfake Epidemic—Here's Why That's a Problem," *Tech Policy Press*, October 10, 2024, https://www.techpolicy.press/minors-are-on-the-frontlines-of-the-sexual-deepfake-epidemic-heres-why-thats-a-problem/.

Zhao, Xia, Limin Wang, Yufei Zhang, Xuming Han, Muhammet Deveci, and Milan Parmar. "A Review of Convolutional Neural Networks in Computer Vision."

Artificial Intelligence Review 57, no. 4 (2024): 99. https://doi.org/10.1007/s10462-024-10721-6.