OUT OF OUR DEPTH WITH DEEP FAKES: HOW THE LAW FAILS VICTIMS OF DEEP FAKE NONCONSENSUAL PORNOGRAPHY

Kate Kobriger[*], Janet Zhang[**], Andrew Quijano[***], and Joyce Guo[****]

Cite as: Kate Kobriger et. al., *Out of Our Depth with Deep Fakes: How the Law Fails Victims of Deep Fake Nonconsensual Pornography*, 28 RICH. J.L. & TECH. 204 (2021).

[*] J.D. Candidate 2023, Columbia Law School; M.P.H. Candidate 2023, Columbia Mailman School of Public Health; B.A. 2014, University of Wisconsin – Madison. The authors would like to thank Alexander Abdo and Professor Steven Bellovin for their guidance, support, and mentorship in writing this article.

[**] B.S. 2021, Columbia Fu Foundation School of Engineering and Applied Science.

[***] B.S. 2019, M.S. 2021, Columbia Fu Foundation School of Engineering and Applied Science.

[****] J.D. Candidate 2024, University of California, Berkeley, School of Law; M.S. 2021, Columbia Fu Foundation School of Engineering and Applied Science; M.S. 2021, Columbia School of Journalism.

ABSTRACT

People have used deep fake technology to generate nonconsensual pornography (NCP) since at least 2017. With technological advances, deep fakes are increasingly easy to create and difficult to identify. This article explores the dearth of both technological and legal recourse for victims of deep fake NCP. It first reviews existing technical solutions for detecting deep fakes, finding that successful deep fake classifiers are often only successful for a short while. As soon as computer scientists publish their work, others build on their discoveries to beat the classification mechanism. Deep fake technology is now so advanced that classification technologies cannot reliably detect them. Because technology cannot help victims of deep fake NCP trace and take down the deep fake content, this article next explores potential avenues for legal redress. Current interpretations of § 230 of the Communications Decency Act immunize websites that host user-contributed content against state civil claims, obstructing victims' attempts to effectuate takedown. While existing scholarship regularly notes that § 230 does not preclude copyright infringement claims, the process of filing a copyright claim is not well-suited to NCP victims and, even if it were, fair use doctrine likely protects websites that host deep fake NCP because of deep fakes' transformative nature and the fact that NCP impacts a market in which most victims do not participate. This article finds that copyright—designed to protect intellectual property, not address sexual violence—is an inappropriate solution to deep fake NCP. This article ultimately concludes that the legislature should revise § 230 of the Communications Decency Act to provide legal recourse to victims of deep fake NCP.

I. INTRODUCTION

[1] Victims[1] of deep fake nonconsensual pornography can suffer devastating impacts on their physical and mental health, employment, and social relationships. However, they have few tools to stop their harassers, remove the damaging content, and find justice. Deep fake nonconsensual pornography (NCP) involves someone creating a sexually explicit image or video using artificially intelligent technology to swap a victim's face onto existing pornographic content without their consent.[2] The accessibility of deep fake creation technology helped fuel this form of abuse.[3]

[2] Section II of this article reviews the technologies that facilitate deep fake creation and explores why deep fakes are difficult to detect and remove. After concluding that technology cannot help victims trace and takedown deep fake NCP, we investigate potential avenues for legal redress. While some argue that the law already addresses the harms of deep fake NCP,[4] Section III discusses how § 230 of the Communications Decency Act (CDA)[5] protects social media companies from legal responsibility for much of the content on their platforms, including deep fake NCP. Scholars

[1] For simplicity, the authors refer to all people whose images are nonconsensually included in pornography as "victims," however we recognize the importance of referring to people affected by sexual violence in terms of their choosing. We referred to the Rape, Abuse, and Incest National Network in reaching this decision. *See Key Terms and Phrases*, RAPE, ABUSE & INCEST NAT'L NETWORK, https://www.rainn.org/articles/key-terms-and-phrases [https://perma.cc/M44L-56MT].

[2] Anokhy Desai, *Explainer: Combatting deepfake porn with the SHIELD Act*, JURIST (Apr. 6, 2021, 12:57 PM), https://www.jurist.org/features/2021/04/06/explainer-combatting-deepfake-porn-with-the-shield-act/ [https://perma.cc/B86N-DX8S].

[3] *Id.*

[4] *See, e.g.*, David Greene, *We Don't Need New Laws for Faked Videos, We Already Have Them*, ELEC. FRONTIER FOUND. (Feb. 13, 2018), https://www.eff.org/deeplinks/2018/02/we-dont-need-new-laws-faked-videos-we-already-have-them [https://perma.cc/HEP7-NU4C].

[5] Communications Decency Act, 47 U.S.C. § 230.

regularly mention one option not blocked by the CDA: copyright infringement. Section IV examines the process for filing a copyright infringement claim, assesses fair use defenses against such claims, and discusses the limitations of using copyright infringement to effectuate deep fake NCP takedowns. In that section, we determine that copyright—designed to protect intellectual property, not to address sexual violence—is an inappropriate solution to deep fake NCP. Ultimately, we conclude that the legislature must update § 230 to secure justice for victims of deep fake NCP.

II. NEW TECHNOLOGIES FACILITATE RAPID AND UNDETECTABLE DEEP FAKE PRODUCTION

A. Introduction to Deep Fakes

[3] Though nonconsensual pornography (NCP) dates back to the 1980s,[6] the recent advent of deep fake technology allowed this form of abuse to rapidly increase. "Deep Learning," the technology behind deep fakes, involves "stitching" one person's face onto another person's body or even creating an entirely new synthetic body, a technique accessible to the general public.[7] In this way, deep fake creators can quickly remake a person's photo into pornography. The technology is so advanced that very little manual action is involved; websites have automatically transformed photographs of clothed people into nude pictures.[8]

[4] NCP dominates the short history of deep fakes. In 2017, Reddit hosted its first deep fake when a user who went by "deepfakes" started

[6] *See, e.g.*, Wood v. Hustler Mag., Inc., 736 F.2d 1084, 1085 (5th Cir. 1984).

[7] *See* Ben Dickson, *What are deepfakes?*, TECHTALKS (Sept. 4, 2020), https://bdtechtalks.com/2020/09/04/what-is-deepfake/ [https://perma.cc/WZX7-JR33].

[8] Drew Harwell, *A shadowy AI service has transformed thousands of women's photos into fake nudes: 'Make fantasy a reality'*, WASH. POST (Oct. 20, 2020, 10:28 AM), https://www.washingtonpost.com/technology/2020/10/20/deep-fake-nudes/ [https://perma.cc/R28S-LUN8].

posting celebrities' faces swapped onto pornography.[9] Since then, filmmakers, meme makers, and even political campaigns have used deep fakes.[10] However, despite their diverse uses, they are mainly used to produce NCP. Sensity AI found that in 2018, between 90% to 95% of online deep fake videos were NCP, and 90% of that NCP specifically featured women.[11]

[5] Differentiating a deep fake from an original photo or video is difficult. In 2020, Facebook launched a contest to develop technology to identify deep fakes, in which over 2,114 participants submitted more than 35,000 models.[12] Facebook tested the deep fake detection models on a dataset of around 100,000 clips developed from over 3,000 Facebook-hired actors.[13] Even the best computer models could only spot the deep fakes

[9] Bryan Clark, *Zuckerberg isn't ready for deepfakes,* TNW (June 28, 2019, 12:55 AM), https://thenextweb.com/news/zuckerberg-isnt-ready-for-the-destructive-nature-of-deepfakes [https://perma.cc/3SQP-FURK].

[10] *E.g.*, Karen Hao & Will Douglas Heaven, *The year deepfakes went mainstream*, MIT TECH. REV. (Dec. 24, 2020), https://www.technologyreview.com/2020/12/24/1015380/best-ai-deepfakes-of-2020/ [https://perma.cc/27UA-T7PK]; Karen Hao, *Members are making deepfakes, and things are getting weird*, MIT TECH. REV. (Aug. 28, 2020), https://www.technologyreview.com/2020/08/28/1007746/ai-deepfakes-memes/ [https://perma.cc/Y5VR-BD44]; Timothy Bella, *Anthony Bourdain's voice was deepfaked in new film. The chef's widow and critics aren't happy*, WASH. POST (July 16, 2021, 12:42 PM), https://www.washingtonpost.com/arts-entertainment/2021/07/16/anthony-bourdain-documentary-ai-deepfake/ [https://perma.cc/V6XS-VT3Y].

[11] Karen Hao, *Deepfake porn is ruining women's lives. Now the law may finally ban it*, MIT TECH. REV. (Feb. 12, 2020), https://www.technologyreview.com/2021/02/12/1018222/deepfake-revenge-porn-coming-ban/ [https://perma.cc/4GPC-7QQD].

[12] Christian Canton Ferrer et al., *Deepfake Detection Challenge Results: An open initiative to advance AI*, FACEBOOK AI (June 12, 2020), https://ai.facebook.com/blog/deepfake-detection-challenge-results-an-open-initiative-to-advance-ai/ [https://perma.cc/59C4-5NRH].

[13] *Id.*

around 65% of the time.[14] More recently, Facebook partnered with Michigan State University to develop a method to recognize fingerprint signatures in deep fake images based on their machine learning model.[15] The research showed promising results, with the team's model parsing method outperforming a baseline model created by randomly shuffling each hyperparameter in the ground-truth set.[16] However, that method still has not been deployed in a real-world setting. The Facebook research lead, Tal Hassner, admitted in a Verge interview that these results might also spur innovation from deep fake creators to outsmart such detection models.[17] Facebook has yet to develop a feasible and scalable solution to effectively identify deep fakes on their platform, underscoring the technical challenge that both distributors and victims face in removing such content.

[6] Technology has now evolved to the point where clicking a few buttons can have life-changing repercussions. NCP is a critical threat to its victims' physical, mental, and financial health. A 2017 study of more than 3,000 victims of NCP observed increased somatic symptoms and decreased

[14] James Vincent, *Facebook contest reveals deepfake detection is still an 'unsolved problem'*, VERGE (June 12, 2020, 11:20 AM), https://www.theverge.com/21289164/facebook-deepfake-detection-challenge-unsolved-problem-ai?fbclid=IwAR3LGa4ZEecJUEE_AnYjUdpA5_N8MRWCy_qgzuO-Arg5UXmEliM5953OgDI [https://perma.cc/QW72-ZZNP].

[15] Vishal Asnani et al., Reverse Engineering of Generative Models: Inferring Model Hyperparameters from Generated Images (June 15, 2021) (unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/2106.07873.pdf [https://perma.cc/FAF8-V27L] (discussing the study by Michigan State on Facebook's newest development regarding fingerprint signature recognition).

[16] Xi Yin & Tal Hassner, *Reverse engineering generative models from a single deepfake image*, FACEBOOK AI (June 16, 2021), https://ai.facebook.com/blog/reverse-engineering-generative-model-from-a-single-deepfake-image/ [https://perma.cc/D46J-HWTA].

[17] James Vincent, *Facebook develops new method to reverse-engineer deepfakes and track their source*, VERGE (June 16, 2021, 12:00 PM), https://www.theverge.com/2021/6/16/22534690/facebook-deepfake-detection-reverse-engineer-ai-model-hyperparameters [https://perma.cc/MW6Q-WTC4].

mental health inventory in victims.[18] Victims of NCP report offline, in-person conduct that threatens their physical wellbeing, such as harassment and stalking as a result of the NCP.[19] The experience almost universally tolls their mental health, with a staggering 51% of victims reporting suicidal thoughts and 93% reporting significant emotional distress resulting from the NCP.[20] The mental health effects are similar to those experienced by rape survivors.[21] NCP also impacts victims' financial wellness by impairing their occupational functioning, which can force them to take time off work or school, cause them to drop out of school, or even result in termination from employment.[22] More than half of victims fear that the NCP will affect their professional advancement even decades into the future.[23]

B. An Overview of Neural Networks That Generate Deep Fakes

[7] Deep fake technology relies on deep neural networks.[24] Neural networks are a subset of machine learning algorithms with structures that

[18] *See* ASIA A. EATON ET AL., CYBER CIVIL RIGHTS INITIATIVE, 2017 NATIONWIDE ONLINE STUDY OF NONCONSENSUAL PORN VICTIMIZATION AND PERPETRATION 24 (2017), https://www.cybercivilrights.org/wp-content/uploads/2017/06/CCRI-2017-Research-Report.pdf [https://perma.cc/WU7D-ARGU].

[19] CYBER CIV. RTS. INITIATIVE, REVENGE PORN STATISTICS 1 (2014), http://www.endrevengeporn.org/main_2013/wp-content/uploads/2014/12/RPStatistics.pdf [https://perma.cc/EGA6-8DNS].

[20] *Id.* at 1–2.

[21] *See* Samantha Bates, *Revenge Porn and Mental Health: A Qualitative Analysis of the Mental Health Effects of Revenge Porn on Female Survivors*, 12 FEMINIST CRIMINOLOGY 22, 33 (2017).

[22] CYBER CIV. RTS. INITIATIVE, *supra* note 19.

[23] *Id.*

[24] *See* Sally Adee, *What Are Deepfakes and How Are They Created?*, IEEE SPECTRUM (Apr. 29, 2020), https://spectrum.ieee.org/what-is-deepfake [https://perma.cc/BH4S-5NEB].

imitate the neurons inside the human brain.[25] They are composed of multiple node layers, including an input layer, hidden layers, and an output layer.[26] Each node is a processing unit within the neural network: it receives data, multiplies it against its assigned weight, and then analyzes and processes it.[27] The result of that analysis is then passed onto the next node, where more analysis is performed.[28] Eventually, each analysis accumulates to produce the final classification result.[29] Together, these layers form a machine learning model that learns and improves accuracy based on training data.[30] For example, to train a deep learning neural network for facial recognition, the programmer would show the model pictures of a given person's face (for our purposes, "Person A") and pictures that are not of Person A's face.[31] The model will then learn the unique features of Person A's face.[32] It may pick up on features like the distance between Person A's eyes or the shape of their nose. As the model trains and learns, it improves the accuracy of categorizing a face as Person A's or not Person A's.[33]

[25] *See Neural Networks*, IBM CLOUD EDUCATION (Aug. 17, 2020), https://www.ibm.com/cloud/learn/neural-networks [https://perma.cc/9NCU-RHXH].

[26] *Id.*

[27] *Id.*

[28] *Id.*

[29] Larry Hardesty, *Explained: Neural Networks*, MIT NEWS (Apr. 14, 2017), https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414 [https://perma.cc/X885-QSU4].

[30] IBM CLOUD EDUCATION, *supra* note 25.

[31] Jane Brownlee, *A Gentle Introduction to Deep Learning for Face Recognition*, MACH. LEARNING MASTERY (July 5, 2019), https://machinelearningmastery.com/introduction-to-deep-learning-for-face-recognition/ [https://perma.cc/68H2-UHC2].

[32] Mei Wang & Weihong Deng, *Deep Face Recognition: A Survey*, 429 NEUROCOMPUTING 215, 216 (2020).

[33] *Id.*

[8] Warren McCullough and Walter Pitts first proposed neural networks in 1944, but the technology remained relatively unknown because of the considerable computing power required to train these networks.[34] With the rise in popularity and the power of graphic processing units (GPUs) computing resources in the 2000s, deep learning has gained acceptance.[35] This rise in popularity allowed deep learning to evolve complex applications, such as creating deep fakes.

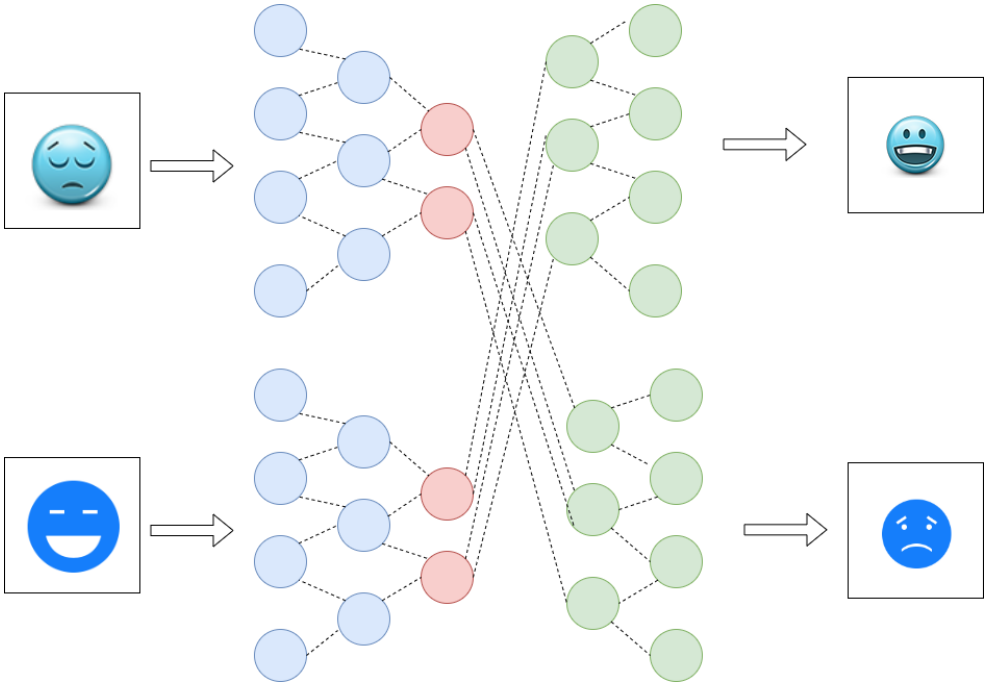

Figure 1: Example of a Facial Reconstruction Autoencoder Neural Network

[34] Hardesty, *supra* note 29, at 1–3.

[35] Tim Dettmers, *Deep Learning in a Nutshell: History and Training*, NVIDIA DEV. (Dec. 16, 2015), https://developer.nvidia.com/blog/deep-learning-nutshell-history-training/ [https://perma.cc/YH9B-Q74P].

[9] Deep fakes rely on two core technologies: (1) autoencoders, and (2) Generative Adversarial Networks (GANs). The above diagram illustrates how an Autoencoder creates a deep fake from two images—in this case, a face swap. An autoencoder is an unsupervised neural network whose purpose is to reconstruct the data it was earlier trained on by identifying key features of the images in the training data.[36] For this reason, autoencoders' input layer (the layer that takes in the images) always has the same number of neurons as their output layer (the layer that outputs the result, in this case, the swapped images).[37] The neural network creates deep fakes by training on the two image sets.[38] A single encoder finds common features between the two sets.[39] Then, there are two separate decoders, each of which is trained on only one image set.[40] Finally, each decoder will use its training to generate an image from the opposite image set, essentially rebuilding a face it has seen before with the data of another face.[41] In plain terms, an autoencoder is like a modern artist who learned to paint only by looking at images of the Mona Lisa (the training data). If that artist were given a description of all of Frida Kahlo's essential features and asked to paint her face, the painting (the output image) would look as if Frida Kahlo had her portrait painted by Leonardo da Vinci. Deep fakes are a seamless merging of two faces.

---

[36] Thanh Thi Nguyen et al., Deep Learning for Deepfakes Creation and Detection: A Survey 1, 3 (Apr. 26, 2021) (unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/1909.11573.pdf [https://perma.cc/X8GL-WSSF].

[37] *See* Arden Dertat, *Applied Deep Learning – Part 3: Autoencoders*, TOWARDS DATA SCI. (Oct. 3, 2017), https://towardsdatascience.com/applied-deep-leaming-part-3-autoencoders-1c083af4d798 [https://perma.cc/X5ZH-6WK8].

[38] *See id.*

[39] Ian Sample, *What are deepfakes – and how can you spot them?*, GUARDIAN (Jan. 13, 2020, 7:48 AM), https://www.theguardian.com/technology/2020/jan/13/what-are-deepfakes-and-how-can-you-spot-them [https://perma.cc/7P9N-HEES].

[40] *Id.*

[41] *See* Daniel Nelson, *What is an Autoencoder?*, UNITE.AI (Sept. 20, 2020), https://www.unite.ai/what-is-an-autoencoder/ [https://perma.cc/6VG5-6LHM].

[10] In the past, creating realistic face manipulations required advanced manual editing skills to ensure the features moved realistically. Today, autoencoders automatically achieve this result, producing flawless face manipulations.[42] Applications like ZAO and FaceApp are implementations of this technology that are easily downloadable and accessible to the public, and they require very little technical skill.[43] Any user who downloads the app can upload and manipulate photos.

[11] Deep fakes fall into four categories, all of which use autoencoders and Generative Adversarial Network:

1. Entire Face Synthesis: The models output an entirely new, non-existent face image.
2. Identity Swap: The models replace one person's face in a photo or video with the likeness of another.
3. Attribute Manipulation: The models modify a person's attributes in a photo or video, such as hair color or skin color.
4. Expression Swap: The models modify the facial expression of a person in a photo or video.[44]

Because deep fake NCP definitionally replaces one person's likeness with another's, this article is primarily concerned with Category 2, Identity Swap.

[42] *See* Ben Dickenson, *What are deepfakes?*, TECHTALKS (Sept. 4, 2020), https://bdtechtalks.com/2020/09/04/what-is-deepfake/ [https://perma.cc/N7XZ-GDY5].

[43] Ruben Tolosana et al., *Deepfakes and Beyond: A Survey of Face Manipulation and Fake Detection*, 64 INFO. FUSION 131, 131 (2020).

[44] *See id.* at 132.

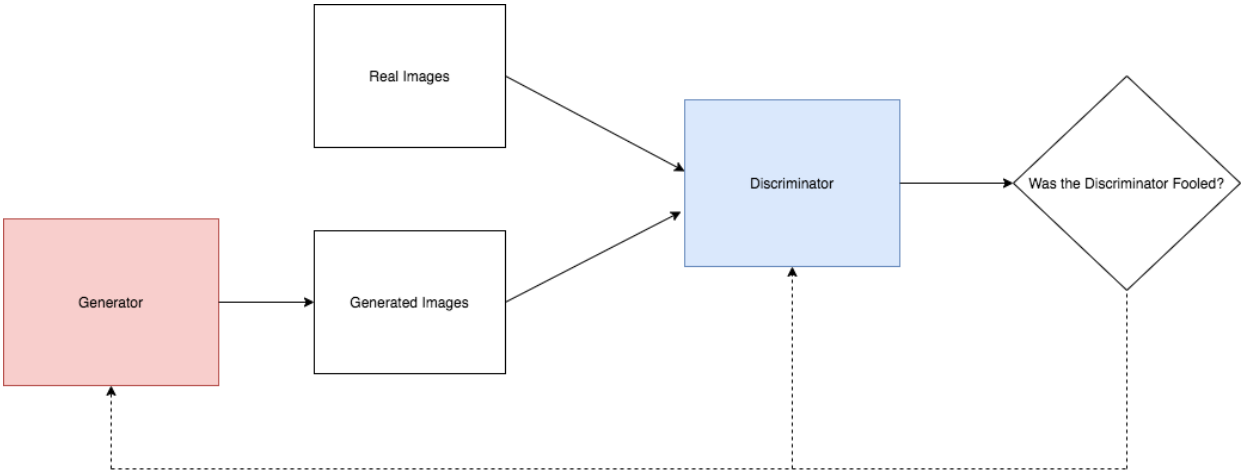

Figure 2: The GAN Model

[12] In addition to an autoencoder, a Generative Adversarial Network (GAN) is needed to create a deep fake. The diagram above depicts a GAN,[45] a neural network that creates a discriminator neural network model, and a generator neural network model. The discriminator model distinguishes whether the incoming data came from training data or the generated model's output.[46] In other words, the discriminator discerns whether an image is real or is a deep fake. The generator attempts to maximize the discriminator's likelihood of incorrectly labeling the input as training data or data coming

[45] Chart created by authors. *See* Ian J. Goodfellow et al., *Generative Adversarial Nets*, ADVANCES NEURAL INFO. PROCESSING SYS. 1 (June 10, 2014), https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf [https://perma.cc/X5HF-FQE7] (defining a GAN as a training framework in which a generative model competes with a discriminative model whereby the generative model produces fake data and real data in a training set and the discriminative model must determine which is the real data).

[46] Kiran Sudhir, *Generative Adversarial Network – History and Overview*, TOWARDS DATA SCI. (June 21, 2017), https://towardsdatascience.com/generative-adversarial-networks-history-and-overview-7effbb713545 [https://perma.cc/RHZ5-SBRA].

from the generator.[47] This process repeats until the discriminator cannot perform better than a random guess (or 50% accuracy) to determine if the data originated from the training data or the generator.[48]

[13] GANs are effective at creating deep fake NCP.[49] Because GANs continually improve, it is almost impossible to distinguish actual content from deep fake content.[50] It is easy to access GAN training data from a social media account, such as a victim's picture. The hardware required to train a GAN is sufficiently cheap for ordinary computer enthusiasts to buy or rent it.[51] There is a significant amount of research into detecting deep fakes to combat the rising issue of distinguishing between deep fakes and real images.[52] However, there is also considerable research on how to

[47] *Id.*

[48] *GAN Training*, GOOGLE DEVS. (Apr. 17, 2019), https://developers.google.com/machine-learning/gan/training [https://perma.cc/99MA-D92P].

[49] Chenxi Wang, *Deepfakes, Revenge Porn, And The Impact On Women*, FORBES (Nov. 1, 2019, 7:39 PM), https://www.forbes.com/sites/chenxiwang/2019/11/01/deepfakes-revenge-porn-and-the-impact-on-women/?sh=60ec5d851f53 [https://perma.cc/HR64-BXVL].

[50] *See* Rob Toews, *Deepfakes Are Going To Wreak Havoc On Society. We Are Not Prepared.*, FORBES (May 25, 2020, 11:54 PM), https://www.forbes.com/sites/robtoews/2020/05/25/deepfakes-are-going-to-wreak-havoc-on-society-we-are-not-prepared/?sh=194d07177494 [https://perma.cc/JMN2-YJGE].

[51] *See* Rajaswa Patil, *Training GANs using Google Colaboratory!*, TOWARDS DATA SCI. (Sept. 16, 2018), https://towardsdatascience.com/training-gans-using-google-colaboratory-f91d4e6f61fe [https://perma.cc/W36B-S9AH].

[52] *See, e.g.*, Pavel Korshunov & Sébastien Marcel, Deepfakes: A New Threat to Face Recognition? Assessment and Detection 1 (Dec. 20, 2018) (unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/1812.08685.pdf [https://perma.cc/SP4M-FFP9] (demonstrating that GAN-generated Deepfake videos are incredibly hard to detect); Xin Yang et al., Exposing Deepfakes Using Inconsistent Head Poses (Nov. 13, 2018) (unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/1811.00661.pdf [https://perma.cc/39D2-XQ5R] (finding that detecting inconsistent 3D head poses could help detect deep fake images or videos).

circumvent deep fake detection systems.[53] The two technologies continue to evolve side-by-side, making legal action one of the few ways to effectively combat deep fake NCP content.

C. Deep Fakes Evade Detection

[14] In 2019, Rössler and colleagues published one of the earliest papers tackling deep fake detection.[54] Their deep fake classifier used neural networks trained on the expansive *Faceforensics++* data set, which comprises thousands of videos of varying resolution, pixel coverage of faces, and a relatively equal proportion of male and female faces.[55] The researchers first investigated whether computer science students could detect deep faked images with the naked eye.[56] Given images from the four models used in the *Faceforensics++* data set, the students' accuracy ranged from 58% to 68.7%.[57] Rössler's deep fake classifier detected deep fakes

[53] *See, e.g.*, Shehzeen Hussain et al., Adversarial Deepfakes: Evaluating Vulnerability of Deepfake Detectors to Adversarial Examples (Nov. 7, 2020) (unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/2002.12749.pdf [https://perma.cc/GFR7-R5FH] (demonstrating it is possible to fool deep neural network-based detectors of deep fake videos).

[54] *See* Felix Juefei-Xu et al., Countering Malicious DeepFakes: Survey, Battleground, and Horizon 12 (Feb. 27, 2021) (unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/2103.00218.pdf [https://perma.cc/L2GG-5422] (providing a table with all papers discussing deep fake detection, listed chronologically).

[55] *See* Andreas Rössler et al., FaceForensics++: Learning to Detect Manipulated Facial Images 2 (Aug. 26, 2019) (unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/1901.08971.pdf [https://perma.cc/3264-YJ8V]; *see also* Justus Thies et al., *Deferred Neural Rendering: Image Synthesis Using Neural Textures*, 38 ACM TRANSACTIONS ON GRAPHICS 1, 9 (2019), https://dl.acm.org/doi/pdf/10.1145/3306346.3323035 [https://perma.cc/C99Q-LRGQ].

[56] Rössler et al., *supra* note 55 at 5.

[57] *Id.* (providing average accuracy rates of about 68.7% for raw videos, about 66.6% for high quality videos, and about 58.7% for low quality videos).

with 95.7% accuracy with high-resolution images and 81% with lower resolution images.[58]

[15] Their success was short-lived. Hussain and colleagues responded to Rössler's research by successfully modifying deep fakes from the *Faceforensics*++ data set to fool Rössler's deep fake classifier at its preprocessing and classification stages.[59] The Rössler classifier used a facial extraction preprocessor to analyze only the faces within input images.[60] Hussain's research targeted the preprocessing stage with techniques such as adding random data from a normal distribution to the input deep fake image and adding extra pixels or shifting them around the image.[61] This is one step that allowed Hussain to fool Rössler's deep fake classifier into falsely identifying deep faked images as likely legitimate 99.05% of the time (under certain circumstances).[62] Even more disturbing, Hussain and colleagues were similarly successful in fooling Rössler's classifier at the classification stage. They sidestepped the need even to access the preprocessing step, which means that any attacker could fool a highly trained classifier model without inside knowledge. Their algorithm reverse engineered Rössler's neural network's use of deep fake image probability, fooling Rössler's deep fake classifier between 84% and 96% of the time.[63]

[16] The Rössler and Hussain dialogue is a notable example of the cat-and-mouse game endemic to deep fake technologies. As soon as someone publishes a deep fake classification method, computer scientists respond

[58] *Id.*

[59] *See* Hussain et al., *supra* note 53 (assuming that an attacker would have complete access to Rössler and colleagues' deep fake detection system, i.e. that it was not a black box to the attacker).

[60] Rössler et al., *supra* note 55, at 4.

[61] Hussain et al., *supra* note 53, at 3351–52.

[62] *Id.* at 3352–53.

[63] *Id.*

with a new way to evade that method. This adversarial model makes consistent and reliable deep fake detection unlikely, if not impossible.

D. Because Deep Fakes Evade Detection and Spread Rapidly, They Require a Legal, Rather Than Technological, Response

[17] Deep fakes have the power not only to fool people but to fool people quickly and widely. This past summer, Chris Ume created a series of deep fake videos of Tom Cruise that garnered more than 11 million views on one social media platform, TikTok, alone.[64] Videos can quickly go viral as they spread across platforms, many of which are international.[65]

[18] Containing viral content is complex because of how quickly it spreads, and because the Internet is constantly archived. Even if a platform removes the content, individual users can screenshot, save, and re-publish information at a speed that outpaces removal moderation.[66] Organized archiving efforts, like the Internet Archive, snapshot, archive, and make publicly available records of the most popular websites on the World Wide Web.[67]

---

[64] Bianca Britton, *Deepfake videos of Tom Cruise went viral. Their creator hopes they boost awareness.*, NBC NEWS (Mar. 5, 2021, 10:02 AM), https://www.nbcnews.com/tech/tech-news/creator-viral-tom-cruise-deepfakes-speaks-rcna356 [https://perma.cc/FNQ2-BRSA].

[65] *E.g.*, Chi Zhang, *How misinformation spreads on WeChat*, COLUM. JOURNALISM REV. (Oct. 30, 2017), https://www.cjr.org/tow_center/wechat-misinformation-china.php [https://perma.cc/Y855-XHX3] (describing how disinformation often spreads quickly on the Chinese social media app, WeChat).

[66] *See, e.g.*, Nick Statt, *Facebook says removing viral COVID-19 misinformation video 'took longer than it should have'*, VERGE (July 28, 2020, 4:56 PM), https://www.theverge.com/2020/7/28/21345674/facebook-covid-19-misinformation-breitbart-news-video-removal-response [https://perma.cc/FAX6-HNX8] (describing Facebook's takedown being delayed for hours).

[67] Kalev Leetaru, *How Much Of The Internet Does The Wayback Machine Really Archive?*, FORBES (Nov. 16, 2015, 9:04 AM), https://www.forbes.com/sites/kalevleetaru/2015/11/16/how-much-of-the-internet-does-the-wayback-machine-really-archive/?sh=327fe45f9446 [https://perma.cc/4UEX-F7RA].

[19] Even if software could identify deep fake pornography with 100% certainty, the rapid spread of content makes it impossible, or at least impractical, to retroactively identify and inform all viewers that they consumed fake content.[68] Further, a viewer who is later informed they consumed deep fake content might not link the notice to the content they viewed, or their perception of a person or an event may be so established by the time they receive the notice that the notice becomes futile.[69] Given these limitations, the only remaining remedy is limiting the distribution of deep fake NCP.

III. COMMUNICATIONS DECENCY ACT § 230 OFFERS NEAR-COMPLETE PROTECTION TO DEEP FAKE DISTRIBUTORS

[20] It is difficult to identify the person who created, or even merely uploaded, deep fake NCP. Users avoiding surveillance find a haven in the Dark Web, a collection of thousands of websites that use tools like Tor and the Invisible Internet Project (I2P).[70] Tor makes the IP address associated with a deep fake upload untraceable, and I2P is a fully encrypted private network layer that encrypts uploaders' data.[71] Recent research found that

[68] *See, e.g.*, Statt, *supra* note 70 (describing Facebook's process of notifying video viewers "who reacted to, commented on, or shared this video, will see messages directing them to authoritative information" after the video had been shared tens of millions of times).

[69] *See* Monica Bulger & Patrick Davison, *The Promises, Challenges and Futures of Media Literacy*, 10 J. MEDIA LITERACY EDUC. 1, 9–10 (2018) (referring to a study where some students stood by opinions they formed after viewing a website even after it was revealed that the website spread misinformation and further failed to recognize how the information was falsified).

[70] Robert W. Gehl, WEAVING THE DARK WEB: LEGITIMACY ON FREENET, TOR, AND I2P 5–6 (2018).

[71] Roger Dingledine et al., *Tor: The Second-Generation Onion Router*, USENIX § 5 (Aug. 2004) https://www.usenix.org/conference/13th-usenix-security-symposium/tor-second-generation-onion-router [https://perma.cc/Q439-UQ2U] (Tor); THE INVISIBLE INTERNET PROJECT, https://geti2p.net/en/ [https://perma.cc/84VY-MZXE] (I2P).

the Dark Web specifically fuels the spread of misinformation,[72] making it an ideal conduit for deep fake NCP. Victims will thus struggle to prove who created or first posted the deep fake NCP, precluding claims against individual uploaders.

[21] However, victims can easily identify which platforms host the deep fake NCP when they are alerted to the deep fake content. Instead of pursuing action against the individual uploader, victims may wish to take action against the platform.[73] Unfortunately, this option is limited by § 230 of the Communications Decency Act (CDA), which broadly immunizes websites that host user-contributed content against state civil claims.[74] As a result, internet platforms escape responsibility for content regulation.

[22] Specifically, § 230 states that computer service providers such as internet platforms are not publishers or speakers of information provided by another "information content provider," regardless of the content.[75] However, a computer service provider becomes an information content provider if they are "responsible, in whole or in part, for the creation or

[72] *See* Sooraj Shah, *Dark web scammers exploit Covid-19 fear and doubt*, BBC (May 19, 2020), https://www.bbc.com/news/business-52577776 [https://perma.cc/9RT4-86T6] (providing an example of how misinformation is being exploited by the Dark Web regarding Covid-19).

[73] *See* Peter A. Halprin et al., *Deepfakes and Insurance Coverage*, N.Y.L.J. (May 7, 2021, 2:50 PM), https://www.law.com/newyorklawjournal/2021/05/07/deepfakes-and-insurance-coverage/ [https://perma.cc/E7RA-VAUG].

[74] Communications Decency Act, 47 U.S.C. § 230(c); *c.f.* § 230(e) (elaborating that contrary state laws are preempted but § 230(e) does not "prevent any State from enforcing any State law that is consistent with this section" and that federal criminal law, property law, and the Electronic Communications Privacy Act are not covered by the immunity provision).

[75] *Id.* §230(c)(1).

development of information."[76] Courts interpret this exception narrowly,[77] so victims arguing that deep fake distributors are information content providers are unlikely to succeed. Therefore, § 230 is significantly responsible for NCP victims' lack of adequate protection.

A. Protections Against Computer-Generated Child Pornography Imply the Same Protections Against Deep Fake Nonconsensual Pornography

[23] Although no deep fake NCP court decisions exist, analogous cases across circuits suggest that platforms hosting deep fake NCP will enjoy the same protections as platforms hosting more traditional NCP.[78] We compare a case in which computer-generated child pornography not involving actual children was considered protected speech with a case where morphed child pornography involving identifiable children was not considered protected speech. In *Ashcroft v. Free Speech Coalition*, stakeholders in the pornography industry successfully challenged the Child Pornography

[76] *Id.* § 230(f)(3).

[77] *See* Fair Hous. Council of San Fernando Valley v. Roommates.com, LLC, 521 F.3d 1157, 1175 (9th Cir. 2008) (holding that websites that provide tools for unlawful conduct are not "information content providers" as long as the tools themselves are neutral, even if the website operators know the tool is being used for unlawful conduct).

[78] *See, e.g.*, United States v. Hotaling, 634 F.3d 725, 729–30 (2d Cir. 2011) (distinguishing the harms caused by computer generated child pornography that implicates identifiable children from that which does not, and finding the former is not protected speech because it implicates the interests of the minors involved); Doe v. Boland, 698 F.3d 877, 879, 884 (6th Cir. 2012) (finding that digitally imposing children's faces onto sexually explicit images implicated the interests of actual children and were therefore indistinguishable from child pornography); United States v. Mecham, 950 F.3d 257, 258 (5th Cir. 2020) (holding that morphed child pornography, which uses the images of real children, is not protected speech even though no contemporaneous pain is inflicted when the image is created); Shoemaker v. Taylor, 730 F.3d 778, 786 (9th Cir. 2013) ("[M]orphed images are like traditional child pornography in that they are records of the harmful sexual exploitation of children.").

Protection Act of 1996 (CPPA) for its overbroad restriction of speech.[79] The Supreme Court relied on the *Miller* standard, which defines obscenity as speech lacking "serious literary, artistic, political, or scientific value,"[80] to conclude that the CPPA's restrictions against what it considered to be child pornography violated the First Amendment.[81]

[24] The CPPA proscribed sexually explicit images that appeared to depict children even when they did not depict children.[82] For example, the CPPA would have banned films that cast adults as teenagers engaging in sexual activity.[83] The same films would not be obscene under the established *Miller* standard because of their artistic or cultural value: youths engaging in sexual activity "has been a theme in art and literature throughout the ages" reflecting our society's "vital interest . . . in the formative years we ourselves once knew."[84] While prior cases permitted states to ban activities "intrinsically related" to the sexual abuse of children,[85] the *Free Speech Coalition* court reasoned that, because the "child

[79] Ashcroft v. Free Speech Coalition, 535 U.S. 234, 245, 262 (2002); *see generally* Child Pornography Prevention Act of 1996, 18 U.S.C. § 2251 (1996) [hereinafter CPPA] (providing categories that restrict speech when involved with child pornography).

[80] *See generally* Miller v. California, 413 U.S. 15, 24 (1973) (establishing the *Miller* standard for obscene material).

[81] *Free Speech Coalition*, 535 U.S. at 246 ("The CPPA, however, extends to images that appear to depict a minor engaging in sexually explicit activity without regard to the *Miller* requirements.").

[82] *See* CPPA § 2256(8)(B) ("[S]uch visual depiction is a digital image, computer image, or computer-generated image that is, or is indistinguishable from, that of a minor engaging in sexually explicit conduct").

[83] *Free Speech Coalition*, 535 U.S. at 246–48.

[84] *Id.* at 246, 248.

[85] *See generally* New York v. Ferber, 458 U.S. 747, 759 (1982) (allowing a state ban on child pornography because production and distribution of child pornography served as a permanent reminder of the child's abuse and created an economic motive to continue its production).

pornography" described in the CPPA does not record any crime or create any victims through its production, there was no direct harm in the speech.[86] The Court acknowledged that "the images can lead to actual instances of child abuse," such as pedophiles using videos to encourage children to engage in sexual activity.[87] Still, it called the causal link "contingent and indirect" and refused to ban products and activities solely because of their potential immoral use.[88]

[25] The Court in *Free Speech Coalition* did not explicitly distinguish between child pornography that uses real children's faces and that which does not. However, that distinction is critical in courts' treatment of NCP. Producing and distributing computer-generated child pornography that does not involve real minors may not directly create victims. Still, when computer-generated pornography uses real children's faces, voices, or bodies, it victimizes those children.[89] *Free Speech Coalition* considered computer-generated child pornography that did not involve minors, so its rationale for protecting pornography creators and platforms is limited.

[26] Indeed, where pornography involves real children, courts firmly protect victims rather than platforms. In *United States v. Hotaling*, the Second Circuit ruled that morphing actual children's faces onto adult bodies performing sexual acts is not protected speech under the First

[86] *Free Speech Coalition*, 535 U.S. at 250, 240 ("As a general rule, pornography can be banned only if obscene, but under *Ferber*, pornography showing minors can be proscribed whether or not the images are obscene under the definition set forth in *Miller* . . . .").

[87] *Id.* at 236.

[88] *Id.*

[89] *See* United States v. Hotaling, 634 F.3d 725, 728, 730 (2d Cir. 2011) ("Unlike the computer generated images in *Free Speech Coalition*, where no actual person's image and reputation were implicated, here we have six identifiable minor females who were at risk of reputational harm and suffered the psychological harm of knowing that their images were exploited and prepared for distribution by a trusted adult.").

Amendment.[90] There, John Hotaling superimposed the images of female minors' heads over images of nude adult females engaging in sexual conduct and labeled the pictures with the minors' first names.[91] The Second Circuit reiterated that the "underlying inquiry is whether an image of child pornography implicates the interests of an actual minor" and followed an Eighth Circuit decision that found minors' interests are implicated when pornography includes their "recognizable face."[92] The *Hotaling* court deftly distinguished *Free Speech Coalition* with a reminder that in *Free Speech Coalition*, "no actual person's image and reputation were implicated, [whereas in *Hotaling*] we have six identifiable minor females who were at risk of reputational harm and suffered the psychological harm of knowing that their images were exploited and prepared for distribution by a trusted adult."[93] The court paid notable attention to the scope of harm in reaching its decision: Not only were minors the only recognizable people in the photos, but the labels including their actual names bolstered the connection between the photos and the minors, increasing the risk of reputational and psychological harm.[94] And while the "harm begins when the images are created,"[95] it does not end there—victims are "haunted for years by the knowledge of [NCP's] continued circulation."[96]

[27] Deep fake NCP causes these same harms.[97] While the *Hotaling* court relied on the government's more specific compelling interest in

[90] *Id.* at 730.

[91] *Id.* at 727.

[92] *Id.* at 729 (citing *United States v. Bach*, 400 F.3d 622 (8th Cir. 2005)).

[93] *Id.* at 730.

[94] *Hotaling*, 634 F.3d at 729.

[95] *Id.* at 730.

[96] *Id.* at 728.

[97] *See, e.g.*, *supra* Section II.A.

protecting minors to outweigh First Amendment considerations,[98] the government has at least a legitimate interest, if not a compelling interest, in protecting public health.[99] Public health includes both minors and adults, and it reaches both mental and physical health,[100] encompassing the full range of deep fake NCP victims and a considerable subset of its harms. This understanding of the law initially appears to open the possibility of a range of negligence and privacy torts against platforms that host deep fake NCP.[101] But even if free speech alone is not enough to shield them, § 230 of the Communications Decency Act broadly immunizes internet platforms from liability, failing to protect both victims of NCP and victims of deep fake NCP.

B. § 230 Gives Websites Broad Immunity for Nonconsensual Pornography and Nonconsensual Deep Fake Pornography

[28] In *Barnes v. Yahoo!, Inc.*, the Ninth Circuit found that § 230 barred a claim for negligent provision of services.[102] Cecilia Barnes's ex-boyfriend created fraudulent profiles on Yahoo containing sexually explicit images of her.[103] Barnes followed Yahoo's policy to request profile removal, and

[98] *Hotaling*, 634 F.3d at 728.

[99] *See, e.g.*, Roman Cath. Diocese v. Cuomo, 141 S. Ct. 63, 67 (2020) (stating that reducing spread of infectious disease is "unquestionably a compelling interest"); Washington v. Glucksberg, 521 U.S. 702, 735 (1997) (finding that state interest in preserving life is "unquestionably important and legitimate").

[100] S. Marshall Williams et. al, *The Role of Public Health in Mental Health Promotion*, MORBIDITY & MORTALITY WKLY. REP. (Sept. 2, 2005), https://www.cdc.gov/mmwr/preview/mmwrhtml/mm5434a1.htm [https://perma.cc/72CQ-JCC6].

[101] *See* Megan Farokhmanesh, *Is it legal to swap someone's face into porn without consent?*, VERGE (Jan. 30, 2018, 2:39 PM), https://www.theverge.com/2018/1/30/16945494/deepfakes-porn-face-swap-legal [https://perma.cc/KW5L-TY33].

[102] Barnes v. Yahoo!, Inc., 570 F.3d 1096 (9th Cir. 2008), *amended by*, Barnes v. Yahoo Inc., 2009 U.S. App. LEXIS 21308, at *1105 (9th Cir. Sept. 28, 2009).

[103] *Barnes*, 570 F.3d at 1098.

further reached out to Yahoo five separate times after receiving no response to her initial request.[104] Yahoo ignored her requests until the night before a local news organization planned to broadcast a report about the incident, when a Yahoo agent finally made a verbal promise to resolve the issue.[105] Nevertheless, Yahoo did not remove the profiles until Barnes filed the lawsuit.[106]

[29] Even though Yahoo's policy explicitly disallowed posting others' content and Yahoo's agent accepted responsibility for removing the NCP,[107] the court found that Yahoo could not be held liable for "negligent undertaking."[108] Because Yahoo did not help develop the unlawful content, it was not a publisher under § 230, and the court held that removing published material is a publisher's duty.[109] Section 230 shielded Yahoo from liability for the negligent undertaking.[110] Whether the content is deep fake NCP or traditional NCP, the outcome would be the same: Yahoo is not legally a publisher of the content and is therefore not legally liable for failing to remove the images. Section 230 leaves victims of deep fake NCP without recourse against distributors even when they explicitly ban such content.

[104] *Id.* at 1098–99.

[105] *Id.*

[106] *Id.*

[107] *Id.* at 1098.

[108] *Barnes*, 570 F.3d at 1105.

[109] *See id.* at 1101–03 (determining Yahoo lacked "publisher" status under 47 U.S.C. § 230(c)(1) as they did not help develop the unlawful content).

[110] *See id.*

C. Federal Trade Commission Actions Are Unlikely to Circumvent § 230

[30] Barnes could have attempted to circumvent § 230 by bringing a complaint to the Federal Trade Commission (FTC). The FTC is a federal agency that, through the Bureau of Consumer Protection, can impose penalties on websites that violate consumer protection laws either in response to consumer complaints, or of its own volition.[111] Under § 5 of the Federal Trade Commission Act, "unfair or deceptive acts or practices in or affecting commerce . . . are . . . declared unlawful."[112] Section 5 enforcement actions thus bifurcate into claims of unfairness and claims of deception. The FTC frequently uses such actions to protect users' privacy,[113] indicating they could fit victims of deep fake NCP. Indeed, both types of claims stand some chance of success, though each must overcome significant obstacles.

---

[111] *See A Brief Overview of the Federal Trade Commission's Investigative, Law Enforcement, and Rulemaking Authority*, FED. TRADE COMM'N (May 2021), https://www.ftc.gov/about-ftc/what-we-do/enforcement-authority?fbclid=IwAR0sE5Z45is8RSzIn5gye-8PRFIq7M-YAIiG2BLnCkOBgJNoU25V7UF5zyw [https://perma.cc/MBH4-KC6S] (citing the Clayton Act and Federal Trade Commission Act).

[112] Federal Trade Commission Act § 5, 15 U.S.C. § 45(a).

[113] *See, e.g.*, Craig Timberg & Tony Romm, *The U.S. government fined the app now known as TikTok for $5.7 million for illegally collecting children's data,* WASH. POST (Feb. 27, 2019), https://www.washingtonpost.com/technology/2019/02/27/us-government-fined-app-now-known-tiktok-million-illegally-collecting-childrens-data/ [https://perma.cc/2DUS-2Q7Y] (explaining that, in response to thousands of complaints from parents of young children, the FTC sued TikTok for illegally collecting names, emails, pictures and locations of kids under 13 years of age; TikTok settled the case for a $5.7 million fine.); *Facebook Settles FTC Charges That It Deceived Consumers By Failing To Keep Privacy Promises*, FED. TRADE COMM'N (Nov. 29, 2011), https://www.ftc.gov/news-events/press-releases/2011/11/facebook-settles-ftc-charges-it-deceived-consumers-failing-keep [https://perma.cc/BL94-4XYZ] [hereinafter *Facebook Settles FTC Charges*] (examining the FTC suing Facebook after it told consumers they could keep their profile private even though third party applications could access private profiles via users' friends' public profiles.).

[31] The circumstances of deep fake NCP will often constitute a *prima facie* § 5 unfairness claim. Unfairness claims require three key elements: "the injury must be (1) substantial, (2) without offsetting benefits, and (3) one that consumers cannot reasonably avoid—as well as the subsidiary role of public policy."[114] In a series of decisions regarding a complaint against LabMD, the FTC and the Eleventh Circuit explored each of these prongs in the context of data privacy.[115] LabMD was a medical laboratory that conducted diagnostic testing using medical specimens and relevant patient information.[116] After an employee broke company policy and downloaded a file-sharing application, LabMD exposed 9,300 consumers' personal information, including names, dates of birth, social security numbers, and health insurance information, and refused a private company's offer to provide remediation services in response to that breach.[117] The FTC found that consumers' injury was substantial, satisfying prong (1): "substantial injury may be demonstrated by a showing of a small amount of harm to a large number of people, as well as a large amount of harm to a small number of people," and "a practice may be unfair if the magnitude of the potential injury is large, even if the likelihood of the injury occurring is low."[118] Finally, while "most cases of unfairness involve economic harm or health and safety risks . . . in extreme cases, subjective types of harm might well

[114] J. Howard Beales, *The FTC's Use of Unfairness Authority: Its Rise, Fall, and Resurrection*, FED. TRADE COMM'N (May 30, 2003), https://www.ftc.gov/public-statements/2003/05/ftcs-use-unfairness-authority-its-rise-fall-and-resurrection [https://perma.cc/XF36-6GRN].

[115] LabMD, Inc. v. FTC, 894 F.3d 1221, 1227–28 (11th Cir. 2018) (declining to disturb the FTC's analysis regarding the requirements of an unfairness claim, but ultimately vacated FTC's cease and desist order demanding LabMD develop a new data security program, which exceeded the scope of congressional intent for § 5); LabMD, Inc., 2014-1 Trade Cases P 78784 (F.T.C.) (2014).

[116] *LabMD, Inc.*, 894 F.3d at 1224.

[117] *Id.* at 1224–25.

[118] F.T.C., LabMD, Inc., Opinion of the Commission, Docket No. 9357, *Federal Trade Commission* (F.T.C.) at 9-10 (Sept. 29, 2016).

be considered as the basis for a finding of unfairness."[119] The FTC cited "'harassing late-night telephone calls' from debt collectors" as an example of subjective harm that would give rise to an unfairness claim.[120] On these grounds, deep fake NCP certainly risks substantial injury sufficient to satisfy the first prong of an unfairness claim. As Section II.A describes, deep fake NCP often results in economic and health harms.[121] Deep fakes' subjective harms align with the subjective harms that gave rise to a claim in the LabMD line of cases. Like LabMD's failure to maintain the security of exposed individuals' sensitive personal information in the form of, for example, herpes or HIV status,[122] deep fake distributors give the impression of exposing parallel information in sexual preferences and experiences. Further, even if most consumers are not victims of deep fake NCP, deep fake NCP risks "a large amount of harm to a small number of people" given the topical and temporal breadth of harms that NCP causes.[123]

[32] The LabMD precedent also supports the finding that consumers cannot reasonably avoid the harms that result from deep fake NCP, satisfying the third prong of the unfairness test. Because patients alternatively did not know their physicians worked with LabMD or "lacked any information about LabMD's data security practices," they "had no opportunity to avoid injuries caused by these practices."[124] In *LabMD, Inc.*, this analysis centered on whether consumers can avoid harm before it occurs and considered whether consumers could mitigate the harm after it

[119] *Id.*

[120] *Id.* at 10.

[121] *See e.g. supra* Section II.A.

[122] *See* LabMD, Inc., *supra* note 118, at 19 (discussing established protections for medical information and tort law's recognition of general "privacy harms that are neither economic nor physical," which become actionable if the matter publicized "(a) would be highly offensive to a reasonable person, and (b) is not of legitimate concern to the public.").

[123] *Id.*

[124] *See* LabMD, Inc., *supra* note 118, at 26.

occurred.[125] Victims of deep fake NCP similarly do not know that anyone with their sensitive information is working on a given platform, and they lack any information about platforms' security practices. Compared to the patients in the LabMD cases, victims of deep fake NCP face a more extreme disadvantage: the LabMD patients entrusted their information to their physicians.[126] In contrast, victims of deep fake NCP do not entrust their image or other information to the deep fake creator. As evidenced in *Barnes v. Yahoo*, consumers have virtually no ability to mitigate the harms of deep fake NCP after it is uploaded,[127] especially because platforms often refuse to assist in a takedown and because deep fake NCP rapidly spreads to other sites.[128] Most importantly, victims of deep fake NCP have no means of avoiding the harms of deep fake NCP before it is shared—they may not even know the NCP exists.[129] Platforms, however, have some ability to screen and moderate content.

[33] Since the solution is likely some form of screening or moderation, victims of deep fake NCP may struggle to win an unfairness claim as they attempt to meet the second prong of § 5 unfairness, requiring that the injury be sustained without offsetting benefits.[130] This prong "is particularly important in cases where the allegedly unfair practice consists of a party's failure to take actions that would prevent" the injuries, and allows benefits as attenuated as "lower costs and then potentially lower prices for consumers."[131] In *LabMD, Inc.*, the record detailed low-cost solutions

[125] *Id.*

[126] *Id.* at 1.

[127] *See* Barnes v. Yahoo!, Inc., 570 F.3d 1096 (9th Cir. 2008), *amended by*, Barnes v. Yahoo Inc., 2009 U.S. App. LEXIS 21308, at *1105 (9th Cir. Sept. 28, 2009); *see supra* Section III.B.

[128] *See supra* Section II.D.

[129] LabMD, Inc., *supra* note 118, at 26–28.

[130] *Id.* at 26.

[131] *Id.*

available to cure LabMD's security deficiencies.[132] However, it is practically impossible to identify deep fake NCP before or after it consistently is posted.[133] Rössler and Hussain's publications demonstrate that any attempt to do so would require ongoing research and development, which could be prohibitively expensive and cannot achieve the desired outcome of quick deep fake detection.[134] There may be a *prima facie* claim of unfairness. Still, a court reviewing this prong on the merits will likely find that attempting to detect deep fake NCP is too costly for the limited benefit of only sometimes detecting and removing it.

[34] Deception claims are more promising. Section 5 deception theory requires three elements: there must be (1) "a representation, omission, or practice that is likely to mislead the consumer[, (2) who is] acting reasonably in the circumstances, [(3)] to the consumer's [material] detriment."[135] The clearest way to satisfy the first prong is through a platform's published policies or terms of service.[136] Though many FTC cases end in settlement, leaving little traditional precedent, the FTC has

[132] *Id.* at 27.

[133] *See supra* Section II.

[134] *See supra* Section II.C (implying that the "cat-and-mouse game endemic to deep fake technologies" would only lead to continuous research and expenses with no end in sight).

[135] FED. TRADE COMM'N, FTC POLICY STATEMENT ON DECEPTION (1983), https://www.ftc.gov/system/files/documents/public_statements/410531/831014deceptionstmt.pdf?fbclid=IwAR1EXD_h3kLs6G_NIEoy13JRcNUYUavfwaapvmes4wJqPEVsG-sdzCuhymY [https://perma.cc/UA77-ZY4J]; *see also* Danielle Keats Citron & Daniel J. Solove, *Privacy Harms*, 102 B.U. L. REV. (forthcoming 2022) (explaining the definition of a "deceptive" act).

[136] FED. TRADE COMM'N, *supra* note 135 ("There may be a concern about the way a product or service is marketed, such as where inaccurate or incomplete information is provided. A failure to perform services promised under a warranty or by contract can also be deceptive.").

historically based claims on platforms' written policies.[137] For example, the FTC brought a case against TikTok for collecting sensitive information about users under the age of 13 despite its policy against accounts for underage users.[138] Similarly, the FTC sued Facebook under a deception theory after Facebook told consumers they could keep their profile private, even though third-party applications could access private profiles via users' friends' public profiles.[139] In the same way, a platform that bans deep fakes or NCP but fails to take action to remove such content makes a misleading representation to its consumers, satisfying prong one. And if there is evidence to satisfy prong one, the remaining elements easily follow. Claims related to deep fake NCP would likely satisfy prong two because consumers can reasonably expect companies to follow their policies,[140] especially under NCP's sensitive circumstances. Moreover, these claims would satisfy prong three because deep fake NCP always effects harms on victims, whether they relate to mental, physical, or financial health and safety.[141] Further, the FTC "considers claims or omissions material if they significantly involve health, safety or other areas with which the reasonable consumer would be concerned."[142] Therefore deep fake NCP cases would almost always contain all three elements necessary to bring a claim under deception theory.

---

[137] *See, e.g.*, FTC, *Facebook Settles FTC Charges*, *supra* note 113 (illustrating, as an example, the policies that Facebook did not uphold which led to the FTC's claims against Facebook).

[138] Timberg & Romm, *supra* note 113.

[139] FTC, *Facebook Settles FTC Charges*, *supra* note 113.

[140] *See, e.g.*, *id.* (explaining that the FTC charged Facebook under a deception theory after Facebook told consumers they could keep their profile private even though third party applications could access private profiles via users' friends' public profiles; *see generally* FED. TRADE COMM'N, *supra* note 135 (explaining the expectations of a reasonable consumer).

[141] *See supra* Section II.A.

[142] FED. TRADE COMM'N, *supra* note 135.

[35] Several platforms currently have stated policies against deep fakes or NCP, placing them within the bounds of a deception claim. For example, Facebook permits users to request deep fake content takedown in its terms of service,[143] and Pornhub provides for removing NCP in its terms of service.[144] Unfortunately, even companies with policies like these fail to remove banned content in a timely manner. For example, in early February 2018, Pornhub responded to celebrity deep fake NCP user reports with a promise to delete the content.[145] After millions of views and a few days without action or further information on their takedown timeline, Pornhub followed through and removed the deep fake NCP.[146] Pornhub's eventual compliance with its policies came in the aftermath of a bevy of public pressure.[147] Indeed, Pornhub and Facebook both omit timeframes for removal from their terms of service or details about their process for reviewing takedown requests and complaints.[148] That lack of specificity could complicate a deception claim because the FTC would have to prove that the policy implied a shorter timeline in such a way that misled reasonable consumers.[149]

---

[143] Monica Bickert, *Enforcing Against Manipulated Media*, META (Jan. 6, 2020), https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/ [https://perma.cc/TPF2-Y2K7].

[144] *Content Removal Request Form*, PORNHUB, https://www.pornhub.com/content-removal [https://perma.cc/X8G4-52QL]; *Terms of Service*, PORNHUB (May 5, 2021), https://www.pornhub.com/information/terms [https://perma.cc/R42T-F5JB].

[145] Damon Beres, *Pornhub continued to host 'deepfake' porn with millions of views, despite promise to ban [UPDATE]*, MASHABLE (Feb. 12, 2018, 9:43 PM), https://mashable.com/article/pornhub-deepfakes-ban-not-working [https://perma.cc/D3NY-6CUR].

[146] *Id.*

[147] *Id.*

[148] *See* Bickert, *supra* note 143; *Terms of Service*, supra note 144.

[149] *See* FED. TRADE COMM'N, *supra* note 135.

[36] A deception claim would be even more complicated if the platform has no written policy about removing deep fake NCP. Without a written policy, there would be little proof the platform made a "representation, omission, or practice" that misled consumers into believing the platform would act.[150] Something like Yahoo's agent's promise in *Barnes v. Yahoo, Inc.*, might qualify as a representation, but simply doing business as a social media platform probably would not.[151] A historical pattern of taking down NCP could rise to the level of a "practice," but would require the platform to act consistently.[152] Finding that historic NCP takedown implies liability for future failures to take down NCP could create a perverse incentive: platforms might avoid moderating any content for fear of developing a "practice" and exposing themselves to high dollar fines.[153] Like the CDA, the FTC Act's blind spots leave victims of deep fake NCP unseen.

D. Courts Offer Limited § 230 Immunity to Websites That Are "Information Content Providers"

[37] As discussed above, § 230's broad immunity is limited to service providers' platforms and does not apply to "information content providers."[154] Indeed, in *Fair Housing Council of San Fernando Valley v. Roommates.com,* the Ninth Circuit held that § 230 did not protect the platform because the platform developed discriminatory content, which transformed Roommates.com into an "information content provider."[155] Before users could search Roommates.com listings, the platform required

[150] *See id.*

[151] *E.g.*, Barnes v. Yahoo!, Inc., 570 F.3d 1096, 1099 (9th Cir. 2008), *amended by*, Barnes v. Yahoo Inc., 2009 U.S. App. LEXIS 21308, at *1105 (9th Cir. Sept. 28, 2009) (stating that Yahoo "would take care of it").

[152] FED. TRADE COMM'N, *supra* note 135.

[153] *Id.*

[154] Fair Hous. Council of San Fernando Valley v. Roommates.com, LLC, 521 F.3d 1157, 1162 (9th Cir. 2008).

[155] *Id.* at 1164, 1169.

them to create profiles that in turn demanded they answer questions about what attributes they preferred their roommates to have, such as a particular sex, sexual orientation, or familial status.[156] The platform hid or displayed potential roommates based on information gleaned from those discriminatory registration questions.[157] Discriminating based on those characteristics violates the Fair Housing Act,[158] and the court held that by designing a registration process that required users to engage in such discrimination, Roommates.com "provided and affirmatively solicited content" that contributed to that discrimination.[159] The Ninth Circuit took care to distinguish the actions of Roommates.com from other common website practices, using search engines and dating websites as examples.[160] It held that search engines are neutral tools that can be used to carry out unlawful searches, and it found that a dating website "that requires users to enter their sex, race, religion, and marital status through drop-down menus, and that provides means for users to search along the same lines" is similarly neutral.[161] Under those circumstances, search engines and dating websites retain CDA immunity because they do not contribute to developing unlawful content, even though users could exploit them for unlawful purposes. On the other hand, Roommates.com developed an inherently discriminatory platform.

[156] *Id.* at 1161.

[157] *Id.* at 1169.

[158] Fair Hous. Council of San Fernando Valley v. Roommate.com, LLC, 2008 U.S. Dist. LEXIS 130811, at *12 (C.D. Cal. Nov. 7, 2008).

[159] *Roommates.com*, 521 F.3d at 1165–67 ("[R]equiring subscribers to answer the questions as a condition of using Roommate's services unlawfully 'cause[s]' subscribers to make a 'statement . . . with respect to the sale or rental of a dwelling that indicates [a] preference, limitation, or discrimination[.]'").

[160] *Id.* at 1169.

[161] *Id.*

E. Nonconsensual Deep Fake Pornography Victims' Claims Are Limited by the Narrow Definition of "Content Provider"

[38] To circumvent § 230 immunity under *Roommates.com*, victims of deep fake NCP might argue that the platform contributed to the creation or development of unlawful content. In *Roommates.com*, the Ninth Circuit interpreted the term "development" as "referring not merely to augmenting the content generally, but to materially contributing to its alleged unlawfulness. In other words, a website helps to develop unlawful content, and thus falls within the exception to [§] 230, if it contributes materially to the alleged illegality of the conduct."[162] This principle was borne out in *Lemmon v. Snap, Inc.*, in which three boys used Snapchat's Speed Filter to capture speeds as high as 123 mph while driving.[163] They crashed, and all three boys died.[164] The boys' families alleged that Snapchat was a "critical cause" of the collision because it developed the Speed Filter, which they argued is content that encourages reckless driving.[165] The district court disagreed and, citing *Roommates.com*, held that the Speed Filter is a "content-neutral tool[]" because it does not contribute to alleged unlawfulness:[166] Applying this rule to deep fake NCP, consider Snapchat's Face Swap Filter, a deep fake filter that replaces the user's face with that of

[162] *Id.* at 1168, 1182; *see also* Dyroff v. Ultimate Software Grp., Inc., 934 F.3d 1093, 1094–95, 1099 (9th Cir. 2019), *cert. denied*, 140 S. Ct. 2761 (2020) (finding that a message board website that facilitated illegal drug sales did not materially contribute to those drug sales even when it recommended illegal drug sales content to specific users).

[163] Lemmon v. Snap, Inc., 440 F. Supp. 3d 1103, 1105 (C.D. Cal. 2020).

[164] *Id.* at 1106.

[165] *Id.*

[166] *Id.* at 1110–11 ("The Speed Filter can be used at low or high speeds, and [Snapchat] does not require any user to Snap a high speed.").

a friend.[167] Like Snapchat's Speed Filter, its Face Swap Filter can be used in various contexts; for instance, users could exploit it to create deep fake NCP. The filter itself augments content agnostically, leaving it to the user to direct the character of its use.[168] Like the search engines and dating profiles that the *Roomates.com* court carefully excluded, Snapchat would enjoy § 230 immunity against claims related to anything produced using its Face Swap Filter because it is just a neutral tool that *could* be used to create unlawful material.[169]

[39] On the other hand, Snapchat would lose its § 230 protection if it materially contributed to developing NCP. For example, consider a filter that allows users to stitch a friend's face onto a naked body performing a sexual act ("Hypothetical Filter"). In contrast to the Face Swap Filter, the Hypothetical Filter would serve no purpose other than to produce pornographic material. For this reason, a court following Roommates.com would likely find that such a filter materially contributes to the illegal acts of creating and disseminating NCP. [170] The filter's creator would not receive § 230 immunity and would almost undoubtedly face civil liability.

---

[167] *See* Michael Nuñez, *Snapchat and TikTok Embrace 'Deepfake' Video Technology Even As Facebook Shuns It*, FORBES (Jan. 8, 2020, 6:30 AM), https://www.forbes.com/sites/mnunez/2020/01/08/snapchat-and-tiktok-embrace-deepfake-video-technology-even-as-facebook-shuns-it/?sh=3b8a61a842c0 [https://perma.cc/PNF9-UCCV] ("Snapchat has a long history of working with this type of AI-driven camera technology, starting with its popular face-swapping camera lens that launched in April 2016."); Josh Constine, *Snapchat lets you Face-Swap with your camera roll, drops paid replays*, TECHCRUNCH (Apr. 21, 2015, 6:41 PM), https://techcrunch.com/2016/04/21/face-swap-camera-roll/ [https://perma.cc/8X5Z-79QW].

[168] *See* Nuñez, *supra* note 167 (explaining how some social media platforms have developed guidelines for use of their technology, but Snapchat has not).

[169] *See Lemmon*, 440 F. Supp. 3d at 1105 (discussing how another Snapchat filter has been ruled a neutral tool, thus making the probability the Face Swap Filter likely would be).

[170] *See* VALERIE C. BRANNON, CONG. RSCH. SERV., LIABILITY FOR CONTENT HOSTS: AN OVERVIEW OF THE COMMUNICATION DECENCY ACT'S SECTION 230 (2019), https://crsreports.congress.gov/product/pdf/LSB/LSB10306 [https://perma.cc/ZY4P-SVVR].

[40] This is a high bar for victims of deep fake NCP. Though people have used services like our Hypothetical Filter in the past,[171] interpretations of § 230 that require the offending platform to materially contribute to the unlawful content are of little use to victims who find themselves on popular services that offer only neutral tools. Under such interpretations, NCP hosted on mainstream platforms such as YouTube, Reddit, or Facebook would likely be immune.

[41] Most courts interpret the law in accord with the Ninth Circuit,[172] but the Supreme Court has not yet ruled on the precise question of how § 230 applies to platforms that host deep fake NCP. Indeed, in 2020 it denied certiorari on a case almost exclusively concerned with § 230 immunity.[173] In a 2020 statement respecting another denial of certiorari, Justice Thomas

---

[171] Samantha Cole, *This Horrifying App Undresses a Photo of Any Woman With a Single Click*, VICE (June 26, 2019, 5:48 PM), https://www.vice.com/en/article/kzm59x/deepnude-app-creates-fake-nudes-of-any-woman [https://perma.cc/B5FE-GKYB] (Such as the discontinued DeepNude application, which "use[d] neural networks to remove clothing from the images of women, making them look realistically nude.").

[172] *E.g.*, Force v. Facebook, Inc., 934 F.3d 53, 64 (2d Cir. 2019); Marshall's Locksmith Serv. Inc. v. Google, LLC, 925 F.3d 1263, 1267 (D.C. Cir. 2019) ("Congress[] inten[ded] to confer broad immunity for the re-publication of third-party content . . . ."); Jane Doe No. 1 v. Backpage.com, LLC, 817 F.3d 12, 18 (1st Cir. 2016) ("There has been near-universal agreement that section 230 should not be construed grudgingly."); Jones v. Dirty World Entm't Recordings LLC, 755 F.3d 398, 408 (6th Cir. 2014) ("[C]lose cases ... must be resolved in favor of immunity . . . .") (quoting Fair Hous. Council v. Roommates.com, LLC, 521 F.3d 1157, 1174 (9th Cir. 2008) (en banc)); Doe v. MySpace, Inc., 528 F.3d 413, 418 (5th Cir. 2008) ("Courts have construed the immunity provisions in § 230 broadly in all cases arising from the publication of user-generated content."); Almeida v. Amazon.com, Inc., 456 F.3d 1316, 1321 (11th Cir. 2006) ("The majority of federal circuits have interpreted [§ 230] to establish broad . . . immunity.") (internal quotes omitted); Carafano v. Metrosplash.com, Inc., 339 F.3d 1119, 1123 (9th Cir. 2003) ("§ 230(c) provides broad immunity for publishing content provided primarily by third parties.") (citation omitted); Zeran v. America Online, Inc., 129 F.3d 327, 330 (4th Cir. 1997) ("Congress recognized the threat that tort-based lawsuits pose to freedom of speech in the new and burgeoning Internet medium.").

[173] Dyroff v. Ultimate Software Grp., Inc., 934 F.3d 1093, 1093–94, 1096 (9th Cir. 2019), *cert. denied*, 140 S. Ct. 2761 (2020).

highlighted that the Supreme Court had not interpreted § 230 in any context since it was passed.[174] Justice Thomas criticized the lower courts' "questionable precedent" for offering platforms "sweeping immunity" and for reading "extra immunity into statutes where it does not belong. . . There are good reasons to question this interpretation."[175] Just months later, Justice Thomas again expressed his dissatisfaction with § 230 in his concurrence in *Biden v. Knight First Amendment Inst. at Columbia Univ.*, where he stated that the federal government "has given digital platforms 'immunity from certain types of suits' [concerning the] content they distribute, but it has not imposed corresponding responsibilities, like nondiscrimination. . . . "[176]

[42] Justice Thomas wrote alone in both *Malwarebytes Inc. v. Enigma Software Group USA, LLC.*, and in *Biden*, leaving the fate of a § 230 case uncertain should it arrive before the Supreme Court.[177] But Justice Thomas is not alone among the judiciary in his distaste for the Circuits' interpretation of the law. The late Chief Judge Katzman of the Second Circuit, in a 2019 partial concurrence and partial dissent, opined that "we have strayed from the path on which Congress set us out" and "caution is warranted before courts extend the CDA's reach any further."[178] He submitted that the CDA should be curtailed, alternatively because "the CDA does not and should not bar relief" when a platform engages in

[174] Malwarebytes, Inc. v. Enigma Software Grp. USA, LLC, 141 S. Ct. 13 (2020) (Thomas, J., concurring).

[175] *Id.* at 13–15.

[176] Biden v. Knight First Amendment Inst. at Columbia Univ., 141 S. Ct. 1220, 1226 (2021) (Thomas, J., concurring) (internal citations omitted).

[177] *See Malwarebytes*, 141 S. Ct. at 13; *see also Biden*, 141 S. Ct. at 1220.

[178] Force v. Facebook, Inc., 934 F.3d 53, 77-80 (2d Cir. 2019) (Katzman, C.J., dissenting).

"matchmaking" of the sort that *Roomates.com* excepted from liability,[179] or because the "duty" that the requested relief assigns to the platform would not require it to act as a publisher.[180] Notably, neither of these alterations to the prevailing interpretation would protect victims of deep fake NCP, since "the failure to remove . . . content, while an important policy concern, is immunized under § 230 as currently written."[181] And for many victims, removal is a primary concern.[182]

[43] Chief Judge Katzman deferred to Congress to update § 230,[183] but Justice Thomas outlined a "[p]aring back [of] the sweeping immunity courts have read into § 230" that is within the Supreme Court's authority.[184] Given the opportunity, the Supreme Court might "consider whether the text of this

[179] *Id.* at 77 (arguing that platforms can be held liable for their "affirmative role" in bringing users together); Fair Hous. Council of San Fernando Valley v. Roommates.com, LLC, 521 F.3d 1157, 1169 (9th Cir. 2008); *see also supra* Section III.D (discussing how courts offer limited § 230 immunity to websites that are "Information Content Providers").

[180] *Force*, 934 F.3d at 77, 83 (2d Cir. 2019) (Katzman, C.J., concurring in part and dissenting in part) ("The duty not to provide material support to terrorism, as applied to Facebook's use of the algorithms, simply requires that Facebook not actively use that material to determine which of its users to connect to each other. It could stop using the algorithms altogether, for instance. Or, short of that, Facebook could modify its algorithms to stop them introducing terrorists to one another. None of this would change any underlying content, nor would it necessarily require courts to assess further the difficult question of whether there is an affirmative obligation to monitor that content.").

[181] *Id.* at 85.

[182] *See infra* Section IV.A (discussing harms of NCP and sought-after remedies).

[183] *Force*, 934 F.3d at 84, 89 ("It therefore may be time for Congress to reconsider the scope of § 230.").

[184] Malwarebytes, Inc. v. Enigma Software Grp. USA, LLC, 141 S. Ct. 13, 18 (2020) (leaving room for legislative action, Justice Thomas held "States and the Federal Government are free to update their liability laws to make them more appropriate for an Internet-driven society."); *see Force*, 934 F.3d at 84, 89 (suggesting Congress reconsider § 230).

increasingly important statute aligns with the current state of immunity enjoyed by Internet platforms."[185]

IV. COPYRIGHT CLAIMS MAY OFFER SOME ASSISTANCE BUT ARE A POOR FIT FOR THE NONCONSENSUAL DEEP FAKE PORNOGRAPHY

[44] Even if future interpretations of the CDA continue to protect platforms against most civil actions, the CDA as written does not shield deep fake distributors from copyright claims.[186] For this reason, many commentators have hinted that copyright may be a source of justice for victims of nonconsensual pornography and nonconsensual deep fakes.[187] However, while copyright infringement remedies initially appear to meet victims' needs, copyright claims are a poor fit for deep fake NCP because the process is inappropriate for many victims, deep fake distributors can defend against copyright claims with relative ease, and even successful suits cannot guarantee victims' privacy in the long-term.

A. At First Glance, Copyright Infringement Remedies Appear Well-Suited to the Needs of Victims of Nonconsensual Pornography

[45] Copyright claims generally require the claimant to have created the image or video, discovered unauthorized distribution, registered the

[185] *Malwarebytes*, 141 S. Ct. at 14.

[186] Erica Souza, *"For His Eyes Only": Why Federal Legislation Is Needed to Combat Revenge Porn*, 23 UCLA WOMEN'S L.J. 101, 115 (2016).

[187] *See, e.g.*, Danielle K. Citron & Robert Chesney, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 CALIF. L. REV. 1753, 1793, 1795 (2019); David Greene, *We Don't Need New Laws for Faked Videos, We Already Have Them*, ELECTR. FRONTIER FOUND. (Feb. 13, 2018), https://www.eff.org/deeplinks/2018/02/we-dont-need-new-laws-faked-videos-we-already-have-them [https://perma.cc/5JDS-DSQM]; Souza, *supra* note 186, at 115–16.

copyright, and made a legal claim against the unauthorized distributor.[188] Copyright infringement remedies offer two primary advantages to NCP victims: distributors must remove the copyrighted material from their platform, and victims may be entitled to money damages for the infringement.[189] Virtually all successful copyright claims result in a court-mandated takedown.[190] While no remedy can make a victim "whole" after NCP dissemination, removing the content from public view is critical to restoring a victim's privacy, as it decreases future harms.[191] In this way, copyright claims bear an advantage over the tort claims that § 230 bars, even if addressing gender-based violence was not copyright's original intent. While a judge could require takedown as part of a tort remedy, torts more often provide money damages.[192]

[46] Of course, financial compensation might be a crucial component of a victim's healing. For example, it might offset the expense of medical care for NCP-related trauma or lost wages related to employment repercussions.[193] Copyright infringement claims might help here, too, but can only include money damages if the claimant demonstrates a slew of additional circumstances, including that the claimant notified the platform of the copyright infringement, the platform had actual knowledge of the

[188] *See* Digital Millennium Copyright Act, Pub. L. No. 105–304, § 512, 112 Stat. at 2860, 2877, 2886 (1998).

[189] *See, e.g.*, *Id.* at 2880, 2882.

[190] *See Id*. at 2884–2885.

[191] *See* EATON ET AL., supra note 18, at 23–24; CYBER CIV. RTS. INITIATIVE, *supra* note 19.

[192] *See* Lindsay Holcomb, *The Role of Torts in the Fight Against Nonconsensual Pornography*, 27 Cardozo J. Equal Rts. & Soc. Just. 261, 280–81 (2021).

[193] *See* CYBER CIV. RTS. INITIATIVE, *supra* note 19.

infringement, and the platform did not act "expeditiously" to remove the material.[194]

[47] A potential disadvantage to copyright infringement remedies is that they usually do not include criminal punishment, which means distributors are not subject to imprisonment.[195] However, imprisoning corporate distributors does not directly serve victims' interests. Imprisonment may be a desirable remedy to protect against physical harm when an individual disseminates NCP, but NCP distributors are often companies whose leaders usually do not target specific people.[196] Imprisoning leaders of those companies could only serve a punitive, not protective, purpose. It might give victims the sense that justice has been served, but a validating decision on another legal claim could achieve the same sentiment.[197]

B. The Copyright Process Is Not Appropriate for All Victims of Nonconsensual Pornography

[48] Copyright remedies may align with NCP victims' needs, but critics widely—and rightly—disparage the path to obtaining those remedies for all

[194] *See generally* Digital Millennium Copyright Act, *supra* note 192, § 512(c)(A)–(C), at 112 Stat. 2880.

[195] 17 U.S.C. § 506(a)(1); 18 U.S.C. § 2319; James Gibson, *Will You Go to Jail for Copyright Infringement?*, THE MEDIA INST. (May 25, 2011), https://www.mediainstitute.org/2011/05/25/will-you-go-to-jail-for-copyright-infringement/ [https://perma.cc/5YXL-LNP7].

[196] *See, e.g.*, State v. VanBuren, 214 A.3d 791, 798 (Vt. 2018); People v. Austin, 155 N.E.3d 439, 451 (Ill. 2019); People of the V.I. v. Roebuck, No. ST-2020-CR-00289, 2021 V.I. LEXIS 5, at *2 (2021) (acknowledging that NCP cases commonly involve former or prospective intimate partners, or those individuals' intimate partners, sharing content without consent).

[197] *See, e.g.*, Amanda L. Cecil, *Taking Back the Internet: Imposing Civil Liability on Interactive Computer Services in an Attempt to Provide an Adequate Remedy to Victims of Nonconsensual Pornography*, 71 WASH. & LEE L. REV. 2513, 2552, 2556 (2014) (arguing that injunctive relief against distributers and removal of the content is the most effective remedy for victims of revenge porn).

kinds of NCP. Importantly, victims can only register a copyright if they are the "author" of the image,[198] offering no hope for victims who, for example, allowed a partner to take photos of them or hire a photographer for a nude photoshoot, later to find those photos posted online. Even victims who did create the image must complete an often-retraumatizing copyright registration process,[199] including filing the sensitive image alongside their personal, identifying information.[200] Victims are effectively required to participate in nonconsensual dissemination by sharing the image with yet another unintended recipient: The United States government, as well as anyone who chooses to review their entry in the Library of Congress public catalog.[201] This step in the copyright process manifests sexist sentiments that victims "did it to themselves" by sharing photos with an intimate partner.[202] Finally, copyright claims may fail altogether if the NCP

[198] 17 U.S.C. § 201(a); *see also* Souza, *supra* note 186, at 115.

[199] Souza, *supra* note 186, at 115–16.

[200] *See Photographs*, U.S. COPYRIGHT OFF., https://www.copyright.gov/registration/photographs/ [https://perma.cc/C5B9-5C4G] (linking to photographs and videos in which a copyright application requires information such as the author's full legal name, citizenship status, and birth date).

[201] *See* Souza, *supra* note 186, at 115–16; *but see* Erica Fink, *To fight revenge porn, I had to copyright my breasts*, CNN BUS. (Apr. 27, 2015, 1:32 PM), https://money.cnn.com/2015/04/26/technology/copyright-boobs-revenge-porn/index.html [https://perma.cc/8SXS-CSCZ] (explaining that the only person to see the pictures is the one processing the copyright and that a special request for relief could be made to prevent the pictures from appearing in the Library of Congress public catalog).

[202] *See* United States v. Jones, 565 U.S. 400, 417 (2012) (Sotomayor, J., concurring) ("[I]t may be necessary to reconsider the premise that an individual has no reasonable expectation of privacy in information voluntarily disclosed to third parties. . . . This approach is ill suited to the digital age, in which people reveal a great deal of information about themselves to third parties in the course of carrying out mundane tasks."); People v. Austin, 155 N.E.3d 439, 451–52 (2019) (refuting societal sentiments that denigrate people who send nude photos stating that "the sharing of a private sexual image in a personal and direct communication with an intended recipient does not demonstrate that the transmission was never intended to remain private").

constitutes "fair use."[203] Even for traditional, non-deep fake NCP, copyright is a bleary avenue.

[49] The same critiques apply to deep fake NCP, but even more forcefully. NCP distribution is "a unique crime fueled by technology,"[204] and the distribution of deep fake NCP is doubly so. Classic NCP cases often involve a person consensually participating in the creation of a pornographic image which is then non-consensually disseminated beyond the agreed-upon recipient.[205] The foundational facts are critically different for deep fake NCP. Because deep fake creators can generate deep fake photos and videos from virtually any image of the victim,[206] identifying the "author" of the image for copyright purposes becomes more difficult.[207] If the victim has an active social media presence, hundreds or even thousands of photos are available to the public. Without context, like a background indicating the location or an outfit indicating the occasion, the victim cannot determine which photo served as the deep fake creator's source.[208] The features replicated by the autoencoder in the deep fake NCP would be present in any photo of the victim's face.[209] Absent evidence of a source photo, the victim has no basis for a copyright claim.

[203] *See infra* Section IV.C (discussing fair use and its four factors).

[204] *Austin*, 155 N.E.3d at 451.

[205] *See, e.g.*, *id.* at 451; CYBER CIV. RTS. INITIATIVE, *supra* note 19, at 1 (stating that 83% of revenge porn victims took nude photos or videos of themselves and shared it with someone else).

[206] *See generally supra* Section II.

[207] *See generally supra* Section II.

[208] Nelson, *supra* note 41.

[209] *See id.*

C. Fair Use Doctrine May Preclude Copyright Infringement Claims Against Nonconsensual Deep Fake Pornography

[50] Even if the victim can somehow ascertain which photo served as the deep fake's source material and prove they authored it, fair use limits copyright's application to deep fakes.[210] Under the Copyright Act of 1976, a primary source of federal copyright law, using another's work is not copyright infringement if that use is "for purposes such as criticism, comment, news reporting, teaching[,] . . . scholarship, or research."[211] In assessing fair use, courts must consider four factors:

> (1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes;
> (2) the nature of the copyrighted work;
> (3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and
> (4) the effect of the use upon the potential market for or value of the copyrighted work.[212]

While courts balance all four factors, only the first and fourth factors apply differently to deep fake NCP than traditional NCP. We, therefore, limit our discussion to those distinguishing factors.

[51] In assessing the first factor, the purpose and character of the use, courts balance whether the goal is to "make a social comment" or "make money."[213] This bodes poorly for victims of deep fake NCP because NCP

[210] Chesney & Citron, *supra* note 187, at 1793–95 (discussing copyright as a claim against deep fakes generally).

[211] 17 U.S.C. § 107.

[212] *Id.*

[213] Original Appalachian Artworks, Inc. v. Topps Chewing Gum, Inc., 642 F. Supp. 1031, 1034 (N.D. Ga. 1986); Pac. & S. Co., Inc. v. Duncan, 744 F.2d 1490, 1496 (11th Cir. 1984).

is rarely commercially motivated.[214] The first factor also encompasses the extent to which the derivative work transforms the original copyrighted work: if it "adds something new, with a further purpose or different character, altering the first [work] with new expression, meaning, or message," it qualifies as transformative.[215] It is thus considered fair use under the first factor.[216]

[52] Unfortunately for victims of deep fake NCP, even the most basic digital alteration is usually enough to qualify as transformative. In *Authors Guild v. Google*, Google uploaded and published digital versions of copyrighted books as part of the Google Books project.[217] The Second Circuit ruled that "Google's making of a digital copy to provide a search function is a transformative use, which augments public knowledge by making available information *about* Plaintiffs' books without providing the public with a substantial substitute for matter protected by the Plaintiffs' copyright interests in the original works or derivatives of them."[218] Like converting paper books to digital, deep fake NCP creators convert source images into new images or videos. And, just as the search function transformed the purpose and use of the copyrighted books, deep fake pornography transforms the purpose and function of the original copyrighted image. A court following *Authors Guild* would find that deep fake NCP is transformative and therefore fair use.

[53] The most challenging fair use factor for victims of deep fake NCP is the fourth factor, the effect of the use on the potential market for the

[214] *See* EATON ET AL., *supra* note 18, at 19 (finding 0% of individuals who non-consensually disseminated pornography did so "to make money off of it").

[215] Campbell v. Acuff-Rose Music, 510 U.S. 569, 579 (1994).

[216] *See id.*

[217] Authors Guild v. Google, Inc., 804 F.3d 202, 207 (2d Cir. 2015).

[218] *Id.*

copyrighted work. Several courts consider this the most crucial factor.[219] However, most people who create deep fake NCP do so out of a desire to harass and intimidate, not to turn a profit.[220] Even if deep fake creators did intend to generate income from the NCP, most victims of NCP could not demonstrate a negative effect on the potential market for the copyrighted work because they never intended to engage in that market— they do not want to sell pornography featuring their likeness.[221]

[54] Victims may wish to approach the potential market from a different angle by more broadly defining who and what make up the potential market. They might argue that deep fake NCP decreases their market value by diminishing their job prospects. Many employers screen candidates' social media,[222] and if a prospective employer discovered a pornographic video of the candidate, they might hesitate to hire them. Similarly, suppose a victim derived monetary value from their likeness through employment as, for example, a model. In that case, they might argue that the deep fake NCP

[219] Original Appalachian Artworks, Inc. v. Topps Chewing Gum, Inc., 642 F. Supp. 1031, 1035 (N.D. Ga. 1986) (citing Triangle Publ'n Inc. v. Knight-Ridder Newspapers, Inc., 626 F.2d 1171, 1177 (5th Cir. 1980)); Marcus v. Shirley Rowley & San Diego Unified Sch. Dist., 695 F.2d 1171, 1177 (9th Cir. 1983); Roger v. Koons, 960 F.2d 301, 311 (2d Cir. 1992).

[220] Joseph M. Sirianni & Arun Vishwanath, *Bad Romance: Exploring the Factors that Influence Revenge Porn Sharing Amongst Romantic Partners*, 6 ONLINE J. COMM. & MEDIA TECH. 42, 60 (2016).

[221] *See* Brian Feldman, *MacArthur Genius Danielle Citron on Deepfakes and the Representative Katie Hill Scandal*, N.Y. MAG: INTELLIGENCER, (Oct. 31, 2019), https://nymag.com/intelligencer/2019/10/danielle-citron-on-the-danger-of-deepfakes-and-revenge-porn.html [https://perma.cc/74SJ-4DYC].

[222] Sarah O'Brien, *Employers check your social media before hiring. Many then find reasons not to offer you a job*, CNBC (Aug. 10, 2018, 9:18 AM), https://www.cnbc.com/2018/08/10/digital-dirt-may-nix-that-job-you-were-counting-on-getting.html [https://perma.cc/7S52-792Z].

decreased the potential market for their likeness.[223] However, while deep fakes are derivative works, where the machine learning model uses data from an original photo to create the deep fake content, courts typically find fair use unless the derivative work usurps the market, or potential market, of the original work, even when the derivative work suppresses the original's value.[224] In *A.V. ex. rel. Vanderhye v. iParadigms, LLC*, iParadigms archived students' homework for future analysis of plagiarism.[225] The students argued that iParadigms' use diminished the market value of their written assignments.[226] The court disagreed, finding that even though iParadigms used the students' work in a commercial context because none of the students planned to sell their homework to be plagiarized, they had no interest in the plagiarism detection market in which iParadigms worked.[227] iParadigms thus did not create a market substitute for the students' work. Unless victims of deep fake NCP intend to create pornography with their likenesses, they will struggle to argue that deep fake NCP is a market substitute under the fourth fair use factor.

[55] The second and third factors of fair use, which pertain to the purpose and character of the use and the nature of the copyrighted work, will most likely fail to protect the copyrighted image. The second factor will fail because the copyrighted image is a photo, which typically makes it more

[223] Jon Bateman, *Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios*, CARNEGIE ENDOWMENT FOR INT'L PEACE, (July 8, 2020), https://carnegieendowment.org/2020/07/08/deepfakes-and-synthetic-media-in-financial-system-assessing-threat-scenarios-pub-82237 [https://perma.cc/8RDS-7GXG].

[224] *See* Harper & Row, Publrs. v. Nation Enters., 471 U.S. 539, 568 (1985) (finding *The Nation* magazine's use of unpublished quotations from a biography of Gerald Ford was not protected by fair use because it took away the publisher's potential market for prepublication excerpts).

[225] A.V. v. iParadigms, LLC, 562 F.3d 630, 634 (4th Cir. 2009).

[226] *Id.* at 640.

[227] *Id.* at 636.

factual than a work of art.[228] The third factor will fail because deep fake NCP is so advanced that a layperson cannot tell it was derived from an original image.[229]

[56] Copyright claims are victims' last defense against deep fake NCP distributors, given § 230's robust protections for platforms like Facebook and Pornhub.[230] But even copyright claims will likely fail given courts' interpretation of the doctrine, since deep fakes are sufficiently transformative and pornography is a distinct enough market product that deep fake NCP would likely find protection under fair use.

D. Copyright Claims May Provide Financial Redress and Nominal Success, but Cannot Eradicate Nonconsensual Deep Fake Pornography in Practice

[57] Finally, victims who successfully navigate the obstacles to registering and enforcing copyrights are unlikely to achieve the ultimate goal of eliminating the NCPs. Although it was not NCP, a 2020 deep fake of Speaker Nancy Pelosi demonstrates the issue: in around four days, a deep fake video of her slurring her speech during a press conference rapidly spread across social media, tallying more than 2 million views.[231] Even for

[228] *See, e.g.*, Salinger v. Random House, Inc., 811 F.2d 90, 99–100 (2d Cir. 1987) (finding that Random House's biography of Salinger infringed Salinger's copyright because it used his original words, which displayed a sufficient degree of Salinger's own creativity as opposed to just facts from his life).

[229] *See, e.g.*, Rogers v. Koons, 960 F.2d 301, 309–10 (2d Cir. 1992) (finding that Koon's sculpture infringed copyright of Roger's photograph because a reasonable lay person could see substantial similarities between the two works of art).

[230] Jesse Lempel, *Combatting Deepfakes through the Right of Publicity*, LAWFARE (Mar. 30, 2018, 8:00 AM), https://www.lawfareblog.com/combatting-deepfakes-through-right-publicity [https://perma.cc/TB44-DQX4].

[231] *Facebook refuses to remove doctored Nancy Pelosi video*, GUARDIAN (Aug. 3, 2020, 1:37 PM), https://www.theguardian.com/us-news/2020/aug/03/facebook-fake-nancy-pelosi-video-false-label [https://perma.cc/FH4G-FFLH].

someone as high profile as Speaker Pelosi, social media platforms did not remove the manipulated content immediately.[232]

[58] Victims of deep fake NCP face many of the same difficulties when attempting a takedown. As discussed above,[233] victims may encounter a daisy chain of reposts on the same site or across many sites. Individuals could download or screenshot the nonconsensual pornography and reupload it months or years later. Actual eradication would require constant vigilance to detect NCP, further complicated for sites on the deep web since they are not searchable through mainstream search engines.[234] Popular websites like Facebook, YouTube, or Reddit might remove NCP as a matter of policy. Other sites may be less responsive—and any deep fake distributors could delay removal in bad faith or simply out of laziness or negligence.[235] Once deep fake pornography is created and posted, it may survive forever.

[232] *Id.* (emphasizing Facebook labeled the video "partly false" and did not immediately remove it); Hannah Denham, *Another fake video of Pelosi goes viral on Facebook*, WASH. POST (Aug. 3, 2020), https://www.washingtonpost.com/technology/2020/08/03/nancy-pelosi-fake-video-facebook/ [https://perma.cc/CU5H-MWT7] ("TikTok, Twitter and YouTube all removed the footage from their platforms after CNN inquired about it Sunday, but it remains on Facebook. The news outlet says the video has been viewed more than 2 million times.").

[233] *See supra* Section II.D.

[234] *See* Cydney Grannan, *What's the Difference Between the Deep Web and the Dark Web?*, ENCYCLOPEDIA BRITANNICA, https://www.britannica.com/story/whats-the-difference-between-the-deep-web-and-the-dark-web [https://perma.cc/WKG5-Z8JS] (finding only 0.03% of Internet sites are discoverable via mainstream search engines).

[235] *See, e.g.*, Barnes v. Yahoo!, Inc., 570 F.3d 1096 (9th Cir. 2008), *amended by*, Barnes v. Yahoo Inc., 2009 U.S. App. LEXIS 21308, at *1105 (9th Cir. Sept. 28, 2009) (showing the platform did not remove nonconsensual pornography despite multiple notices and a verbal promise to do so, until news media planned a broadcast about it); *see* Denham, *supra* note 232 (highlighting TikTok, Twitter, and YouTube only removed the Nancy Pelosi deep fake following a news media inquiry about it).

V. CONCLUSION

[59] It is time to abandon backdoor theories of liability and address the distributor's role in nonconsensual pornography directly. For years, advocates labored to pass "revenge porn" statutes criminalizing individuals' participation in creating and disseminating nonconsensual pornography.[236] But even as more states adopt such laws, distributors escape responsibility for their complicity through the Communications Decency Act's strong shield. Until Congress amends § 230 of the Communications Decency Act, victims cannot rely on tort law theories,[237] and FTC actions are unlikely to fill the gaps. Further, victims of deep fake NCP have little hope in copyright claims because they require registering their deep fake NCP as copyrighted material. It is difficult to determine which images generated the deep NCP, and defendants can argue the deep fake NCP is fair use. An amendment to § 230 will help victims stem the spread of deep fake NCP by allowing them to target distributors with a more complete and more apt range of civil actions.

[236] *See, e.g.*, *48 States + DC + One Territory Now Have Revenge Porn Laws*, CYBER CIV. RTS. INITIATIVE, http://www.cybercivilrights.org/revenge-porn-laws [https://perma.cc/LGJ6-BTTV].

[237] S*ee* Souza, *supra* note 186, at 114–15.