The Technical, Legal, and Ethical Landscape of Deepfake Pornography

Amelia O'Halloran

April 2021

Advisor
Timothy Edgar

Second Reader
James Tompkin

Submitted in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science with honors at Brown University

Abstract

Deepfakes are doctored videos created by manipulating existing video, photo, and/or audio in order to depict people saying and doing things that never actually occurred – analogous to Photoshop, but built to create realistic videos instead of still images. This novel technology, popularized in the media by depicting doctored videos of politicians speaking or deceased celebrities appearing in new projects, has potentially abusive and exploitative use cases. While deepfake videos seen in news outlets typically surround politics or celebrities, the vast majority of deepfakes online are pornographic. This paper describes the societal harms caused by deepfake pornography. After considering these societal impacts, I contend that victims of deepfake pornography need paths to justice through civil or criminal liability, and online platforms should assist the efforts to support victims by banning deepfake pornography. In order to understand how these efforts can be accomplished, I first detail the technical aspects of deepfake creation and detection, then enumerate the potential civil liability methods for victims, how state and federal law can be altered to criminalize deepfake pornography, and steps that platforms can take to remove deepfake pornography from their websites.

Acknowledgements

Thank you to my thesis advisor, Timothy Edgar, for offering expert insights into privacy law and technology policy. When I was a freshman, I attended a lecture you gave on surveillance and privacy that ignited my interest in technology policy. Two years later, my focus on digital sexual privacy originated in your Computers, Freedom, and Privacy class. We formally started this thesis a year ago, but it really originated almost four years ago in your lecture when you showed me what it meant to work at the intersection of technology and civil liberties.

Thank you to my reader, James Tompkin, for lending invaluable technical expertise to this paper.

Thank you to my Computer Science advisor, Tom Doeppner, for allowing me to achieve the Brown information session mantra of truly being the architect of my own education. My experience in the Computer Science department was exactly what I hoped it would be because of your advising and support.

Thank you to Danielle Citron for setting the standard for sexual privacy research. My passion for this topic stemmed from reading your papers, and I consider you to be my biggest role model in the fight for sexual privacy.

Thank you to the people who made my Computer Science experience so warm, supportive, and truly fun: Madelyn Adams, Daniel Adkins, Henry Peebles-Capin, and William Jurayj. I am indescribably grateful for all of the time we've spent together.

Finally, thank you to my parents for the constant encouragement and for asking about my thesis on every phone call. Thank you to my mom for sending me every single deepfake-related news item she could find online, and to my dad for reading drafts and offering suggestions. Thank you to my two older brothers, Jack and Cormac, for motivating all of my academic endeavors. Everything I have ever done is because you two showed me how. Thank you.

Contents

"The intensity and complexity of life... have rendered necessary some retreat from the world... so that solitude and privacy have become more essential to the individual; but modern enterprise and invention have, through invasions upon his [or her] privacy, subjected him [or her] to mental pain and distress, far greater than could be inflicted by mere bodily injury."

Samuel D. Warren and Louis Brandeis, The Right to Privacy (1890).

"No woman can call herself free who does not own and control her own body."

Margaret Sanger (1919).

Introduction

Deepfake videos are a futuristic science-fiction reality come to life: they can resurrect deceased actors for new scenes, incorrectly convince the public that a political candidate claimed something damaging to his or her campaign, display emergency officials announcing a fake impending missile strike or natural disaster,[1] falsely depict a celebrity engaging in disreputable behavior, or portray any other unlikely situation a deepfake creator can imagine. While deepfake coverage in the media typically surrounds deepfakes of politicians or celebrities, the original usage of deepfakes was nonconsensually inserting victims' faces into pornographic videos.[2] Now, the vast majority of deepfake videos online are pornographic – a shocking 96 percent.[3] Deepfake pornography is an act of digital sexual manipulation that allows creators to invade the privacy and autonomy of anyone – most often, an everyday woman that the creator knows in real life – and reduce her to a mere sexual object. As Mary Anne Franks, notable expert on the intersection of civil rights and technology, stated: "If you were the worst misogynist in the world, this technology would allow you to accomplish whatever you wanted."[4]

The intention of this paper is to bridge the knowledge gap between technologists and policymakers regarding the issue of deepfake pornography. This knowledge gap exists in all technology policy spheres; in 2012, the Secretary of Homeland Security, in charge of assessing and protecting against national cyberthreats, declared at a conference: "Don't laugh, but I just don't use email at all." In 2013, United States Supreme Court Justice Elena Kagan revealed the same was true for eight of the nine Supreme Court Justices.[5] Those deciding what is legal, pertinent, or threatening in the digital sphere are not always educated on the technical aspects. As Supreme Court Justice Warren Brandeis once said, "The greatest dangers to liberty lurk in insidious encroachment by men of zeal, well-meaning but without understanding."[6] Two branches of "well-meaning but without understanding" may exist here: technologists are often

[1] Danielle Citron and Robert Chesney, "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security," *California Law Review 107,* no. 6 (December 1, 2019): 1776.

[2] Samantha Cole, "Deepfakes Were Created As a Way to Own Women's Bodies—We Can't Forget That," *Vice*, June 18, 2018, https://www.vice.com/en/article/nekqmd/deepfake-porn-origins-sexism-reddit-v25n2.

[3] Henry Ajder et al., "The State of Deepfakes: Landscape, Threats, and Impact," *Deeptrace,* September 2019, 27.

[4] Drew Harwell, "Fake-Porn Videos Are Being Weaponized to Harass and Humiliate Women: 'Everybody Is a Potential Target,'" *The Washington Post,* December 30, 2018, https://www.washingtonpost.com/technology/2018/12/30/fake-porn-videos-are-being-weaponized-harass-humiliate-women-everybody-is-potential-target/.

[5] P.W. Singer and Allan Friedman, "Cybersecurity and Cyberwar: What Everyone Needs to Know," Oxford University Press 2014, 5.

[6] Olmstead v. United States, 277 U.S. 438 (1928).

well-meaning in their novel technical innovations but lacking an understanding of the societal harms. Meanwhile, policymakers may attempt to ameliorate the harms of the internet, but without an understanding of the technical aspects, proposed laws may have unintended consequences or lack the coverage of future implications for the technology. This paper attempts to mitigate part of this danger by closing this knowledge gap in technology policy related to deepfake pornography.

While there have been papers related specifically to deepfakes – notably, Citron and Chesney's paper "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security," these papers usually focus on deepfakes in general, or deepfakes in relation to national security. This paper is intended to focus on deepfake pornography, the most widely distributed form of deepfakes on the internet. The first section of this paper details the harmful societal impacts of deepfake pornography, attempting to explain the monumental abuse and exploitation inherent to digital sexual privacy invasions. In order to truly understand deepfake pornography, I offer a technical description of deepfakes in general, which makes up the second section of this paper. This section, along with the technical detection methods offered in the third section, are meant to be read and understood by technologists and policymakers alike. The fourth section consists of potential legal and policy methods that can be used in conjunction with the aforementioned technical detections to offer recourse for victims of deepfake pornography.

The atrocities of deepfake pornographic videos do not lie solely in the fact that they are deepfakes, or the fact that they are pornographic. If done correctly and consensually, deepfakes can be used for benevolent, or at least recreational, purposes: education, art, film, games, and self-autonomy can all be enhanced with forms of deepfake videos, and people with certain forms of paralysis, such as ALS, can create videos allowing them to speak with their own voice.[7] This paper is also not a commentary on the pornography industry as a whole, nor is it claiming that pornography should be banned. It is not deepfakes nor pornography alone that constitute such an atrocious abuse, but the combination of both that creates a massively exploitative digital abuse. Deepfake pornography eliminates the sexual autonomy and sexual privacy of the victim and engenders fear and misogyny online, creating a new avenue for digital sexual abuse.

[7] Citron and Chesney, "Deep Fakes: A Looming Challenge," 1769-1771.

1 Societal Impacts

As a famous actress, Kristen Bell is accustomed to performing as someone other than herself. She speaks lines written for her, displays emotions as explained by writers, and walks and moves her body as directed. However, even though her body performs as another person, she is still cognizant of these actions and personally chooses to perform them. The power Bell holds in her consent to perform was shattered when her husband informed her of the existence of pornographic videos online that appeared to feature Bell herself. Unlike the popular phenomenon of "celebrity sex tapes" dotting the underside of the internet, Bell's video was not one in which she actually participated. Rather, her face was superimposed onto another woman's body in a pornographic video, making it appear as though she was the one participating. When Bell found out, she was incredulous: "this is my face, it belongs to me," she argued, "I don't consent."[8] Not only did she not consent, but she wasn't even aware of the video's existence until after other people had already viewed it, stripping her not only of autonomy but also of knowledge of her own likeness's actions. Unfortunately, this is not uncommon for celebrities. Actors from the silver screen now find themselves in sordid scenes, doctored but still realistic and exploitative.

Bell's non-consensual and falsified deepfake pornographic video is unfortunately not uncommon; deepfake-dedicated websites display video after video featuring various celebrities' likenesses in intimate situations. However, among the most disturbing facts surrounding deepfake pornography is the suprisingly low proportion of celebrities featured in these videos. An anonymous poll undertaken in a Telegram channel that serves as a central hub for creation, discussion, and distribution of deepfake videos asks: "Who are you interested to undress in the first place?" Only sixteen percent of respondents chose "Stars, celebrities, actresses, singers"; eight percent chose "Models and beauties from the internet"; seven percent chose "Just different girls from the internet"; 63% of respondents chose "Familiar girls, whom I know in real life."[9] The majority of these videos are made not to satisfy a lurid fantasy about a movie star one has lusted after (which is still exploitative and non-consensual), but rather as a form of virtual sexual coercion and abuse that allows people to virtually undress and take advantage of women they know. This exploitation of non-celebrity, everyday people was the case for Noelle Martin, who was only eighteen when she found doctored pornographic images of herself online after performing a reverse-image search of her face. Someone had photoshopped Martin's face, taken from Facebook photos uploaded by Martin or her friends, onto pictures of someone else's

[8] Vox. *The Most Urgent Threat of Deepfakes Isn't Politics, 2020.* https://www.youtube.com/watch?v=hHHCrf2-x6w&feature=emb_logo.

[9] Vincent, James. "Deepfake Bots on Telegram Make the Work of Creating Fake Nudes Dangerously Easy." *The Verge*, October 20, 2020. https://www.theverge.com/2020/10/20/21519322/deepfake-fake-nudes-telegram-bot-deepnude-sensity-report.

body. Using the same dataset of Facebook photos, these photographs eventually led to doctored pornographic videos featuring Martin. After being informed of the deepfake videos in an email, Martin recalled watching the videos: "I watched as my eyes connected with the camera, as my own mouth moved. It was convincing, even to me."[10] Martin's connection with her body was altered as she watched herself performing falsified sexual actions and was convinced of their veracity. These videos not only took away Martin's power over how others perceived her body, but also her own understanding and acceptance of her bodily autonomy. She describes the numbing, helpless, and persistent feeling of the exploitation: "I would contact the site administrators and request to get things deleted. Sometimes they would delete the material and then a month later, it would reappear. Sometimes they simply didn't respond... One site host told me they would only delete the material if I sent them nude photos of myself within 24 hours. I had to watch powerless as they traded content with my face doctored into it... I had spent my entire adult life watching helplessly as my image was used against me by men that I had never given permission to of any kind."[11] Noelle stated that these videos not only affected her mental state upon viewing, but she also has anxieties surrounding the lasting effects of these videos on her life, saying that she fears that her future children and future partner will see these videos.[12]

The stories of Kristen Bell and Noelle Martin are not isolated by any means. As machine learning capabilities grow, their abuses follow suit. In September 2019, Deeptrace Labs produced a report detailing the current state of deepfake videos, finding that the total number of deepfake videos online at the time was 14,678 – almost double the number found in December 2018.[13] This swift growth represents the exponential increase of accessibility to sophisticated machine learning algorithms (which I discuss in the Deepfake Creation section of this paper) in order to create deepfake videos. Though people may have heard of deepfake videos in the context of doctored political videos[14] or resurrection of deceased celebrities in films,[15] the vast majority are used for a different purpose; of the deepfake videos online, 96% feature pornographic content.[16] The majority of these videos are found on deepfake-specific pornographic websites.[17] Despite these websites being relatively new, with the earliest deepfake-specific

[10] Daniella Scott, "Deepfake Porn Nearly Ruined My Life," *Elle*, June 2, 2020, https://www.elle.com/uk/life-and-culture/a30748079/deepfake-porn/.

[11] Ibid.

[12] Vox, The Most Urgent Threat of Deepfakes Isn't Politics.

[13] Ajder et al., "The State of Deepfakes," 1.

[14] Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman, "Synthesizing Obama: Learning Lip Sync from Audio," *ACM Transactions on Graphics* 36, no. 4 (July 20, 2017), https://doi.org/10.1145/3072959.3073640.

[15] "Deepfake Tech Drastically Improves CGI Princess Leia in Rogue One — GeekTyrant," accessed March 22, 2021, https://geektyrant.com/news/deepfake-tech-drastically-improves-cgi-princess-leia-in-rouge-one.

[16] Ajder et al., "The State of Deepfakes," 1.

[17] Id., 6.

pornographic website being registered in February 2018, the total number of video views across the top four dedicated deepfake pornography websites as of September 2019 was over 134 million.[18]

It's important to analyze the environment that spawned deepfake pornography. As the digital landscape evolves in unpredictable and immeasurable ways, the situations of Kristen Bell, Noelle Martin, and other deepfake pornography victims fall into a larger phenenomem of digital sexual exploitation. New extensions of the digital world beget corresponding extensions of digital sexual exploitation; some examples of digital sexual exploitation include sextortion,[19] video voyeurism,[20] up-skirt videos,[21] revenge porn,[22] and tracking victims of intimate partner violence relationships.[23] All of these abuses serve to violate the sexual privacy of those exploited. Sexual privacy is defined by Danielle Citron as the "social norms that govern access to, and information about, individuals' intimate lives."[24] Just as Warren and Brandeis argued for the right to privacy in 1890 in order to establish a right to be left alone and to control one's own thoughts, communications, and sentiments,[25] sexual privacy is further intended to give individuals the ability to decide "the extent to which others have access to and information about people's naked bodies (notably the parts of the body associated with sex and gender); their sexual desires, fantasies, and thoughts; communications related to their sex, sexuality, and gender; and intimate activities."[26] Noelle Martin's description of her fears surrounding her future children and future partner seeing the doctored videos are a common effect of sexual privacy invasions. Martin's fears are based on the idea that the invasion of sexual privacy causes a loss of the ability to form one's own self-identity, as the victim feels that others already have a perception of them based on the privacy invasion. Citron writes, "Being able to reveal one's naked body, gender identity, or sexual orientation at the pace and in the way of one's choosing is crucial to identity formation. When the revelation of people's sexuality or gender is out of their hands at pivotal moments, it can shatter their sense of self."[27] Citron continues by explaining that deepfake pornographic videos "can lead to a single aspect of one's self eclipsing one's personhood. Sex organs and sexuality stand in for the whole of one's identity, without the self-determined boundaries that protect us from being simplified and judged out of context."[28] While this thiev-

[18] Id., 1.

[19] Danielle Keats Citron, "Sexual Privacy," *Yale Law Journal* 128, no. 7 (May 1, 2019), https://digitalcommons.law.yale.edu/ylj/vol128/iss7/2, 1915.

[20] Id., 1909.

[21] Id., 1914.

[22] Id., 1917.

[23] Sam Havron et al., "Clinical Computer Security for Victims of Intimate Partner Violence."

[24] Citron, "Sexual Privacy," 1874.

[25] Samuel D. Warren and Louis D. Brandeis, "The Right to Privacy," *Harvard Law Review 4,* no. 5 (1890): 193–220, https://doi.org/10.2307/1321160.

[26] Citron, "Sexual Privacy," 1880.

[27] Id., 1884.

[28] Id., 1925.

ery of autonomy can be viewed as unethical through a plethora of moral lenses, there may be none as obvious as the Kantian Humanity Formula, an ethical proposal stating that people should never act in such a way that we treat humanity, whether in ourselves or in others, as a means only but always as an end in itself.[29] Utilizing people's photos as a means of self-pleasure is an obvious example of treating a person as a means instead of an end. To understand a person as an end within herself, we should be focusing on her humanity and autonomy, as opposed to using her as a self-indulgent tool.

The ability to form one's own identity is especially relevant in cases of intimate love. People develop relationships of love and intimacy through a process of mutual sharing, self-disclosure, and vulnerability.[30] Strengthened sexual privacy allows people to feel secure in the privacy of their sexual life and behaviors so that any disclosure is chosen by oneself in a safe and caring manner. It allows people to "'giv[e]... themselves over' to each other physically and 'be what they truly are—at least as bodies—intensely and together'" in order to express love and care.[31] Once one's sexual privacy has been invaded, it can sometimes wreak lasting effects on one's ability to trust loved ones or friends, impacting the ability to develop intimate relations.[32]

For many women, sexual privacy and the ability to create one's own identity is a crucial aspect of building one's career. Sexual and pornographic content online may cause women to suffer stigmatization in the workforce, lose their jobs, and have difficulty finding new ones[33] – even if the videos are fake. Citron cites a Microsoft study that states that "nearly eighty percent of employers use search results to make decisions about candidates, and in around seventy percent of cases, those results have a negative impact. As another study explained, employers often decline to interview or hire people because their search results featured 'unsuitable photos.'"[34] Explicit images posted online without consent can have negative impacts on one's job search, and doctored explicit images or videos take that a step further, as the one whose likeness appears in the photo may not be aware of its existence, and therefore unaware that her job prospects are being hindered by false images or videos.

In 1869, John Stuart Mill argued for a new system of equality for both sexes, writing that the subordination of one sex to another is one of the chief hindrances to human improvement.[35] More than 150 years ago, he understood

[29]Robert Johnson and Adam Cureton, "Kant's Moral Philosophy," The Stanford Encyclopedia of Philosophy (Spring 2021 Edition), ed. Edward N. Zalta, https://plato.stanford.edu/archives/spr2021/entries/kant-moral/

[30]Citron, "Sexual Privacy," 1875, 1889.

[31]Id., 1888.

[32]Id., 1924.

[33]Id., 1875.

[34]Citron, "Sexual Privacy," 1927.

[35]John Stuart Mill, "The Subjection of Women," in *Feminism: The Essential Historical*

that the imbalance of power between the sexes was dangerous for the progress of humanity. In order to fully understand the harm caused by deepfake pornography, we must place it in this context: deepfake pornography is an explicit tool to subordinate women and create a dangerous power dynamic. According to the Deeptrace Labs report, 100% of the pornographic deepfake videos found in their September 2019 report featured videos created from female images being imposed onto the videos, as opposed to zero percent of videos being created from images of male faces (the report does not break down gender further than male/female).[36] This solidifies what most would assume about any new trend in exploitative online behavior: the victims of these videos are basically exclusively female. While the report does not break down victims by race, it does note that the nationality of the subjects in deepfake pornography videos are about two-thirds from Western countries and about a third from non-Western countries, including South Koreans specifically making up about 25% of the victims.

The overwhelming gender disparity in the victims of this abuse shows that deepfake pornography follows in the footsteps of other invasions of sexual privacy in that it serves as a tool for men to subject and exploit women. The case of deepfake pornography differs from other invasions of sexual privacy, as the victim may not ever be aware that her privacy was invaded, may have never participated in any intimate actions with her abuser, and may have never even met or interacted with her abuser. The dangers of sexual privacy invasions can be amplified when the victim does not know that she was abused; no measures can be taken to mitigate harm when the harm is unknown to the victim. On the other hand, if the victim is aware of the videos, she has to deal with the immense pain of seeing her body playing a role that she never agreed to, participating in intimate situations that she never consented to, and having her sexuality publicized in a way she may not be comfortable with. Deepfake pornography can "transform rape threats into a terrifying virtual reality."[37] The loss of bodily autonomy and the dissociative experience of seeing one's body performing false actions in a realistic video is an indescribable form of abuse. Throughout this paper, as I discuss the potential methods of recourse for victims of deepfake pornography, it is important to return to the idea that these victims are almost exclusively female; thus, working towards a safer digital video environment is not only a matter of general privacy, but specifically a matter of gender equality.

Writings, ed. Miriam Schneir (New York: Vintage Books, 1972), 166.

[36] Ajder et al., "The State of Deepfakes," 2.

[37] Citron and Chesney, "Deep Fakes: A Looming Challenge," 1773.

2 Deepfake Creation

2.1 Overview of Deepfakes

The earliest instances of video editing were performed in the 1950s with razor blades and videotapes. Video editors would splice and piece together frames of a videotape recording, a process that required "perfect eyesight, a sharp razor blade, and a lot of guts."[38] Now, video editing is much more sophisticated and accessible; the process to create an accurate, lifelike doctored video only requires access to a computer and the ability to gather photos and videos of subjects. The underlying mechanism for deepfake creation consists of deep learning models such as autoencoders and Generative Adversarial Networks.[39] I detail the technical aspects of deepfakes below.

Neural Networks

Neural networks are intended to parallel some functions of the human brain in a more simplified manner.[40] As computational models that have the ability to adapt or learn, to generalize, and/or to cluster or organize data,[41] neural networks are intended to learn and train themselves without having to be programmed explicitly. In the case of deepfake models, a neural network may be used to recognize the face of a subject, given hundreds or thousands of photos of that subject. In a non-neural network, the creator of the deepfake could potentially code the explicit dimensions of the subjects face, telling the computer that if a face has a nose length of exactly 5.6 centimeters, a distance between the eyes of 6.3 centimeters, and red hair, then the face belongs to the subject. However, a seemingly-infinite stream of if-statements covering all dimensions of a human face would be time-consuming and challenging to code, so neural networks perform that work for you, learning from photos of the subject to understand how to categorize unseen faces into subject and non-subject.

Neural networks are essential for deepfake videos because they give computer models the ability to learn about the faces of the subjects, which allows them to perform the face-swapping function to place one subject's face onto

[38] Dr Robert G. Nulph, "Edit Suite: Once Upon a Time: The History of Videotape Editing," *Videomaker (blog),* July 1, 1997, https://www.videomaker.com/article/2896-edit-suite-once-upon-a-time-the-history-of-videotape-editing/.

[39] Thanh Thi Nguyen et al., "Deep Learning for Deepfakes Creation and Detection: A Survey," *ArXiv:1909.11573 [Cs, Eess],* July 28, 2020, http://arxiv.org/abs/1909.11573.

[40] Chris Woodford, "How Neural Networks Work - A Simple Introduction," Explain that Stuff, June 17, 2020, http://www.explainthatstuff.com/introduction-to-neural-networks.html.

[41] Ben Krose and Patrick van der Smagt, "An Introduction to Neural Networks," November 1996, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.18.493.

another's body without having to do a frame-by-frame copy-and-paste of one face on top of the other.

Encoder-Decoder Pairs

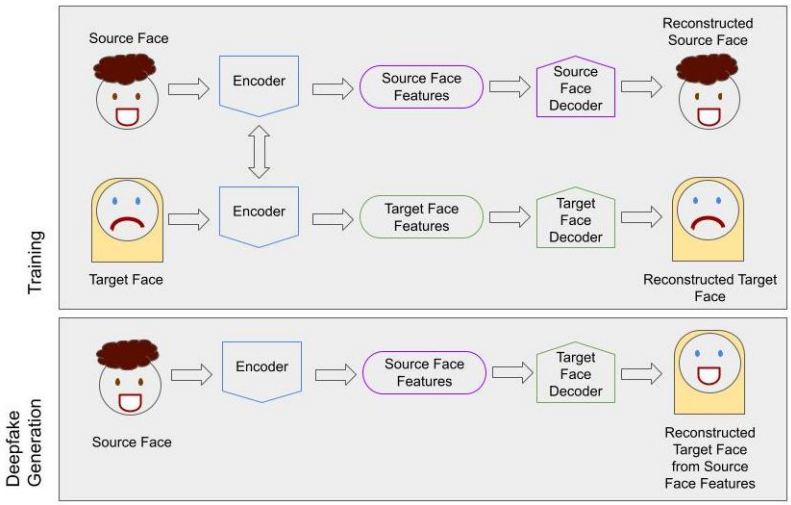

Figure 1: A representation of utilizing encoder-decoder pairs to generate deepfakes.

The original method of creating deepfakes was proposed by a Reddit user using autoencoder-decoder pairing structure.[42] In this method, the autoencoder extracts important features of face images, such as the nose, eyes, and mouth. The decoder uses this information to reconstruct the face image based on the important features. In order to superimpose a target face onto a source face, an encoder-decoder pair is required for both faces where both pairs have the same encoder network. This common encoder finds and learns the similarity among the features in the two sets of face images. Thus, if the encoded feature set of the source face is connected with the decoder from the target face, this reconstructs the target face using the features from the source face. Figure 1 shows an example of this process, where the features from the source face are decoded with the target face's decoder, creating a reconstructed version of the target face with the features (in this example, a smile instead of a frown) of the source face. This creates a video of what appears to be the target subject performing actions that were actually performed by the source subject, making a deepfake of the target subject.

[42] Nguyen et al., "Deep Learning for Deepfakes Creation and Detection."

Generative Adversarial Networks

The advent of Generative Adversarial Networks (GANs), invented by Google researcher Ian Goodfellow, allowed for an immense increase in realism in deepfakes.[43] GANs essentially pit two neural networks against each other: one as a "generator" and the other as a "discriminator." To understand the purpose, I begin with an analogy. Say that Jenny makes her living by creating counterfeit paintings and selling them as if they were real. She has learned the techniques of the great painters and uses them to create very realistic-looking matches. However, she knows that if she gets caught just once creating a counterfeit, then she won't be able to continue this career. Thus, instead of just counting on her own judgement, she hires Desiree, an experienced counterfeit-spotter, to assist in her con. Now, Jenny paints many counterfeit paintings, but not all pass Desiree's judgement; the ones that Desiree accepts are strong enough matches to pass in the eyes of art collectors as well, and the ones that Desiree rejects allows Jenny to learn more about which techniques to avoid and which to enhance. Desiree, as a discriminator, makes Jenny's generated output stronger and more accurate. Thus, generators and discriminators work together to create accurate outputs – in the context of this paper, generators create doctored frames or videos, and discriminators only allow the most realistic ones to be output, causing a more lifelike deepfake video.

[43] Ian J. Goodfellow et al., "Generative Adversarial Networks," *ArXiv:1406.2661 [Cs, Stat],* June 10, 2014, http://arxiv.org/abs/1406.2661.

2.2 Coding a Deepfake Model

Coding a deepfake model can be accomplished by almost anyone, especially if you have some experience using the command line. In order to code a deepfake, one needs access to a Graphics Processing Unit (GPU) to cut down on processing time. Attempting to code a deepfake on a non-GPU computer may take weeks, whereas a GPU only takes up to a few days. To code a deepfake, I used an online repository created by the user 'deepfakes' on GitHub.[44] After cloning the GitHub repository so I can utilize it for my own project, the four stages of coding a deepfake are gathering data, extracting faces, training a computer model, and converting a video using our trained model. I provide a brief overview of each stage of the instructions provided by the user 'deepfakes' below:

1. **Data gathering.** We need to gather data on the appearances of both the subject whose face will be transplanted onto the video and the subject whose face will be replaced. For an accurate deepfake, thousands of photos are required from a variety of angles, lighting, and facial expressions. Instead of trying to search for a dataset of thousands of photos, I used publicly available videos split into separate frames. A few videos per subject (especially where the only face in the video was the subject's) was enough for thousands or tens of thousands of frames.
2. **Extracting faces.** Once we have accumulated our photos, we need to focus on the subjects' faces. User 'deepfakes' provides a command to extract the faces from the photos. Running this command will take a few hours per subject. This allows the model to learn information about the individual faces of each subject by ignoring the background and/or any other people in the photos.
3. **Training a computer model.** Next, we create a model that contains information about the faces of both of our subjects in order to learn how to swap between the two. This will likely take 12-48 hours if using a GPU and weeks if using a CPU.
4. **Converting a video.** Now that we have a functioning model, all we have to do is convert the frames of a video from depicting subject A's face to depicting subject B's face. Once we have the individual frames with faces swapped, we can combine the frames to create a full video. We have now created a deepfake video consisting of frames converted from displaying faces of subject A to faces of subject B on subject A's body.

James Vincent, self-proclaimed non-technical reporter, explains that this process is accessible to even those without a technical background, though it re-

[44] Deepfakes, *Deepfakes/Faceswap,* 2021, https://github.com/deepfakes/faceswap.

quires some patience.[45] There are even applications that will perform these tasks for you to create a deepfake,[46] though applications' outputs will likely be less realistic than those coded using the steps above, as their process is likely simplified.

[45] James Vincent, "I Learned to Make a Lip-Syncing Deepfake in Just a Few Hours (and You Can, Too)," *The Verge*, September 9, 2020, https://www.theverge.com/21428653/lip-sync-ai-deepfake-wav2lip-code-how-to.

[46] Natasha Lomas, "Deepfake Video App Reface Is Just Getting Started on Shapeshifting Selfie Culture — TechCrunch," accessed March 22, 2021, https://techcrunch.com/2020/08/17/deepfake-video-app-reface-is-just-getting-started-on-shapeshifting-selfie-culture/.

3 Technical Detection of Deepfakes

3.1 Overview

The prohibition of deepfake pornography depends on an essential initial step: detection. How can we detect deepfakes when they have become so sophisticated that the human eye often cannot discriminate between real and fake? Luckily, machine learning detection is a more effective classification mechanism than human detection, as I will demonstrate in the following discussion of detection methods. As discussed, this paper is intended to bridge the gap between policymakers and technologists. Thus, the purpose of this section is to not only inform the technologically-inclined on the variety of deepfake detection methods already in existence (in order to augment a platform's detection systems or to provide a foundation for further detection research, for example), it is also intended to inform those who do not have a background in Deep Learning. In order to demonstrate the current state of deepfake detection, I offer summaries of eleven research papers, comprising the most state-of-the-art detection methods in categorizing deepfake videos.

In 2018, Korshunov and Marcel stated that most of the research being done on deepfakes was about advancing the creation of the face-swapping technology, trying to make them more and more realistic. The detection of deepfakes took a backseat to the augmentation of their creation. Keep in mind that the original usage of deepfakes was deepfake pornography, so the fact that research was primarily performed on augmentation of creation as opposed to detection speaks to the lack of socially-conscious judgement applied by many deepfake researchers. However, in 2018, researchers started to work on databases and detection methods.[47] Research is still being done on the generation of more sophisticated and realistic deepfakes, so detection methods are fighting an uphill battle. This relationship may seem like Sisyphus's endless stone-rolling attempts, or even an arms race between creation and detection – and it is. As detection gets more sophisticated, creators know what to look for and how to enhance their videos. Progress in detection forces video generators to be more creative in their realism, and progress in creation realism requires detectors to be more creative in how to spot minute discrepancies or tampering artifacts. Although this paper offered an overview of deepfake creation above, I focus more on detection than creation in order to contribute to what I hope to be the future of deepfake research emphasizing detection and socially just deepfakes.

Monumental increases in deepfake sophistication has led to a division between two generations of deepfakes. First generation deepfakes are characterized

[47] Pavel Korshunov and Sebastien Marcel, "DeepFakes: A New Threat to Face Recognition? Assessment and Detection," *ArXiv:1812.08685 [Cs],* December 20, 2018, http://arxiv.org/abs/1812.08685.

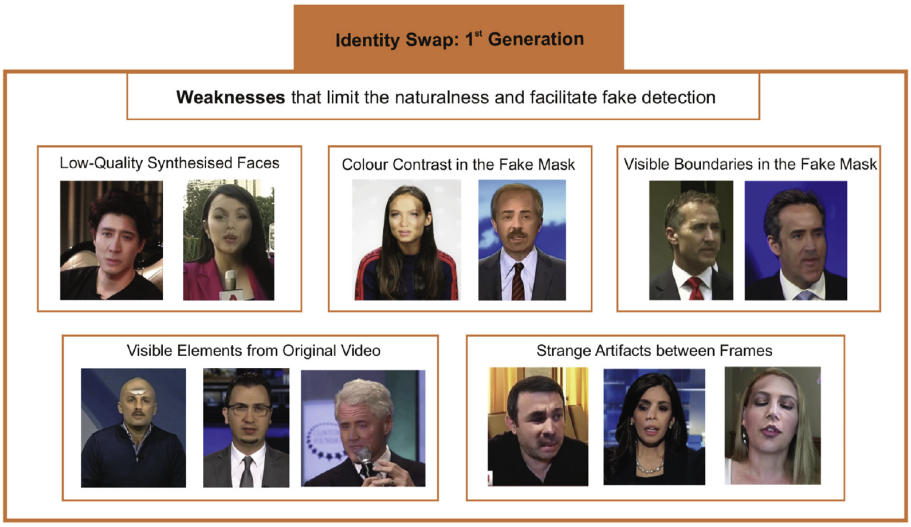

Figure 2: Issues present in first generation deepfakes

by low-quality synthesized faces, different color contrast among the synthesized fake mask and the skin of the original face, visible boundaries of the face mask, visible facial elements from the original video, low pose variations, and strange artifacts among sequential frames – as shown in Figure 2[48] – whereas second generation deepfakes have generally improved in sophistication and realness as new deepfakes often consider "different acquisition scenarios (i.e., indoors and outdoors), light conditions (i.e., day, night, etc.), distances from the person to the camera, and pose variations."[49] Thus, as I will soon summarize a variety of technical detection papers, it is important to note that some of these methods were trained and tested on first generation deepfakes, making them useful for detecting low-quality deepfakes but not more modern, high-quality ones. For example, once blinking was used as an artifact to detect deepfakes, the next generation of synthesis techniques incorporated blinking into their systems, making this detection technique less effective.[50] Therefore, some of the methods below may not detect the current generation of deepfakes, and a substantial number of them may not detect future generations of deepfakes as the realm of detection and creation is constantly evolving.

Aside from the constant improvement of deepfake generation technologies, another challenge of deepfake detection is inherent to the nature of videos. The

[48] Ruben Tolosana et al., "DeepFakes Evolution: Analysis of Facial Regions and Fake Detection Performance," *ArXiv:2004.07532 [Cs],* July 2, 2020, http://arxiv.org/abs/2004.07532.
[49] Ibid.
[50] Shruti Agarwal et al., "Protecting World Leaders Against Deep Fakes," 2019.

precursor to deepfake videos was photoshopped (or otherwise doctored) images, so image forensics have had much longer to gain sophistication and accuracy. However, traditional image forensics often do not translate accurately to videos due to video compression (which often occurs in the storage and posting of videos in order to save space) that strongly degrades the data, making it more difficult to find evidence of tampering.[51]

Finally, before diving into the technicalities of the papers listed below, I find it important to note that of the eleven papers listed, only four mention the dangers of deepfakes in pornography, whereas six mention the danger of deepfakes in politics. Thus, even those interested in the detection of deepfakes and creating literature surrounding this important issue are misunderstanding (or miscommunicating) the proportional reach of the issue. Papers that discuss the political impacts of deepfakes while leaving out the sexual privacy issues are focusing on an important but small percentage of deepfakes and disregarding the most common deepfake exploitation. This may distort the research, as the nature of the activities performed in political deepfakes versus pornographic deepfakes differ – for example, while measuring lip synching may be an effective method for campaign speeches, it may not be as relevant for pornographic videos. This is a clear example of biased datasets that may have real consequences on pornographic deepfake detection, once again demonstrating that sexual privacy, and especially the sexual autonomy of women, is not brought to the forefront in the issue of deepfakes. To those researching deepfake generation or deepfake detection, I urge you to think not only of the political implications, but also of the crucial importance that your work has on victims of sexual privacy invasions.

[51] Darius Afchar et al., "MesoNet: A Compact Facial Video Forgery Detection Network," in *2018 IEEE International Workshop on Information Forensics and Security (WIFS),* 2018, 1–7, https://doi.org/10.1109/WIFS.2018.8630761.

3.2 Measurements of Technical Detection Methods

These methods are difficult to compare in a concrete manner, since the metrics used in each paper vary. Some papers use Accuracy, some use Area Under the Curve, and some use Equal Error Rate. Not only do the papers use different metrics, but they were also tested at varying points in the progression of deepfake generation techniques, adding to comparison difficulties. Therefore, I present these summaries as a means of showing the promising breadth of technical detection possibilities, not in order to rank their accuracies. In order to understand the meanings of the various metrics and other aspects of the following papers, I offer a general definition of various terms that come up in detection methods:

- True Positive (TP): Items that are categorized as true and are true (in this context, items that are categorized as deepfakes and are, in fact, deepfakes).
  - True Positive Rate (TPR) is the percentage of true items that are marked as true – in mathematical terms, $\frac{TP}{TP+FN}$ and in our context, the percentage of deepfakes that are indeed marked as deepfakes.
- True Negative (TN): Items that are categorized as false and are false (in this context, items that are categorized as real videos and are, in fact, real videos).
- False Positive (FP): Items that are categorized as true but are false (in this context, items that are categorized as deepfakes but are actually real videos).
  - False Positive Rate (FPR) is the percentage of negative items that are marked as positive – in mathematical terms, $\frac{FP}{FP+TN}$ and in our context, the percentage of real videos that are inaccurately marked as deepfakes.
- False Negative (FN): Items that are categorized as false but are true (in this context, items that are categorized as real videos but are actually deepfakes).
- **Accuracy:** The percentage of correct categorizations; mathematically, $\frac{TP+TN}{TP+TN+FP+FN}$.

– **Area Under the Curve (AUC):** Models can be optimized to create the highest TPR or lowest FPR. For example, if we had a model that marked all videos automatically as deepfakes (a trivial and non-helpful model), using this model on a dataset of 99 deepfakes and one real video would output a TPR of 100% and an FPR of 100%, which may seem misleadingly promising if the researcher is only focusing on TPR. Instead of looking at only one, we can create a curve with axes consisting of the TPR and FPR; the area under this curve correlates to the correctness of the categorization when not optimized solely for one or the other, but rather at many different points of optimization decisions. Figure 3[52] shows an example of this curve, displaying points on the curve at two different decision thresholds. Thus, the higher the AUC, the better the model performs, as it performs closest to a maximal TPR and a minimal FPR.

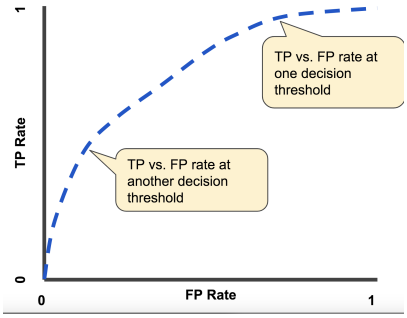

Figure 3

– **Equal Error Rate (EER):** The EER is a metric that again utilizes the AUC curve shown in Figure 3. The EER is the point at the intersection of the AUC curve and the line found by plotting all points where TPR = FPR. The EER represents the optimal performance of the model, as it is the point when the TPR and FPR are both optimized together (TPR being maximized and FPR being minimized) while still equal. Thus, the lower the EER, the better the model performs, as it means a low FPR and a high TPR in the method's most optimal form. The EER point is shown in Figure 4.[53]

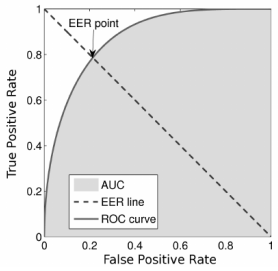

Figure 4

– Convolutional Neural Network (CNN): "A deep learning neural network designed for processing structured arrays of data such as images... Con-

[52] "Classification: ROC Curve and AUC — Machine Learning Crash Course," Google Developers, accessed March 22, 2021, https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc.

[53] "Dynamic Score Combination: A Supervised and Unsupervised Score Combination Method - Scientific Figure," ResearchGate, https://www.researchgate.net/figure/An-example-of-a-ROC-curve-its-AUC-and-its-EER_fig1_225180361.

volutional neural networks are very good at picking up on patterns in the input image, such as lines, gradients, circles, or even eyes and faces... Convolutional neural networks contain many convolutional layers stacked on top of each other, each one capable of recognizing more sophisticated shapes. With three or four convolutional layers it is possible to recognize handwritten digits and with 25 layers it is possible to distinguish human faces."[54]

[54] "Convolutional Neural Network," DeepAI, May 17, 2019, https://deepai.org/machine-learning-glossary-and-terms/convolutional-neural-network.

3.3 Technical Detection Methods

The following list is a subset of methods offered by Tolosana et al. in their 2020 paper.[55] This list is intended to demonstrate the breadth of current detection methods.

1. Korshunov and Marcel (2018) present two methods of deepfake detection: an audio-visual approach and an image-based approach. The audio-visual approach detects inconsistency between lip movements and speech in order to "distinguish genuine video, where lip movement and speech are synchronized, from tampered video." This lip syncing-based algorithm had low accuracy, as the generators of the videos were able to generate facial expressions that match audio speech quite accurately. The second method implemented three different types of image-based systems, the most successful of which used a combination of 129 metrics of image quality, including signal to noise ratio, specularity, blurriness, and so on, to create an average image quality score for 20 frames per video. These image-based approaches are capable of effectively detecting deepfake videos, with the method having a reasonably high accuracy of detecting low-quality deepfake videos with a 3.33% EER and high-quality deepfake videos with an 8.97% EER.[56]

2. Matern et al. (2019) use a visual features-based method. They trained models on both real and deepfake videos to search for discrepancies in missing reflections and details in eyes and teeth, then tested those models on two publicly available datasets with AUCs of 85.1% and 83.8%.[57]

3. Whereas the previous deepfake detection models have relied on visual or audio-visual inconsistencies based on artifacts from the facial synthesis, Yang et al. (2019) propose a new approach: 3D head pose estimation. The 3D head pose "corresponds to the rotation and translation of the world coordinates to the corresponding camera coordinates," and can be used to estimate facial landmarks. In the process of creating deepfakes, "the landmark locations of fake faces often deviate from those of the original faces," so 3D head pose estimation can demonstrate these deviations. Their most successful AUC outcome for five 3D head pose feature classifications was 89% for frames and 97.4% for videos.[58]

[55] List compiled by Ruben Tolosana et al., "Deepfakes and beyond: A Survey of Face Manipulation and Fake Detection — Elsevier Enhanced Reader," December 2020, https://doi.org/10.1016/j.inffus.2020.06.014.

[56] Korshunov and Marcel, "DeepFakes: A New Threat to Face Recognition?"

[57] Falko Matern, Christian Riess, and Marc Stamminger, "Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations," in *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW),* 2019, 83–92, https://doi.org/10.1109/WACVW.2019.00020.

[58] Xin Yang, Yuezun Li, and Siwei Lyu, "Exposing Deep Fakes Using Inconsistent Head Poses," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and*

4. Agarwal et al. (2019) has created a model that distinguishes individuals based on the presence and strength of twenty specific action units or facial motion features such as inner brow raise, outer brow raise, lip tighten, nose wrinkle, cheek raise, head rotation, mouth stretch and more. Each individual is given a "motion signature" based on the frequency and strength of these twenty features, then any deepfake created was tested against this motion signature to see the extent to which they match up. Motion signatures that do not match were classified as fake. This study had a very promising AUC of up to 99% for face-swap deepfakes, however it can only be used in instances where we have a robust set of videos of the individual in order to create his or her motion signature. This study was performed on Barack Obama, Donald Trump, Bernie Sanders, Hillary Clinton, and Elizabeth Warren; it could only be performed accurately on other politicians or celebrities with sufficiently available video data.[59]

5. Jung et al. (2020) analyzes blinking patterns of subjects in a video to determine the veracity of the video.[60] The paper contends that this method achieved accurate results in 7 of the 8 videos tested (87.5%), however I note that blinking is an aspect of deepfake generation that has, after coming under scrutiny in detection methods, been considered more heavily in the creation phase – meaning that more and more deepfake generators ensure that the blinking patterns are stable in their videos. Thus, this method may not be a long-lasting one.

6. Afchar et al. (2018) propose two methods for detection, both at the mesoscopic (intermediate stage between microscopic and macroscopic) level. The methods vary slightly, but generally alternate layers of convolution (filtering) and pooling (reducing the amount of parameters) before being input into a classification network. These methods have achieved accuracies of 96.9% and 98.4% of deepfake video classification.[61]

7. Zhou et al. (2018) uses a two-stream CNN to classify deepfakes: the first stream is a face classification stream that is meant to detect tampering artifacts (such as blurred areas on foreheads, sharp edges near lips, et cetera) in the faces of subjects, and the second stream focuses on features such as camera characteristics and local noise residuals. This two-pronged approach achieved a strong AUC of 92.7% after testing on a different dataset than the model was trained on.[62]

Signal Processing (ICASSP), 2019, 8261–65, https://doi.org/10.1109/ICASSP.2019.8683164.

[59] Shruti Agarwal et al., "Protecting World Leaders Against Deep Fakes."

[60] Tackhyun Jung, Sangwon Kim, and Keecheon Kim, "DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern," *IEEE Access* 8 (2020), https://doi.org/10.1109/ACCESS.2020.2988660.

[61] Darius Afchar et al., "MesoNet: A Compact Facial Video Forgery Detection Network."

[62] Peng Zhou et al., "Two-Stream Neural Networks for Tampered Face Detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, https://doi.org/10.1109/CVPRW.2017.229.

8. Nguyen et al. (2019) propose the use of a capsule network as opposed to a CNN for deepfake detection. Whereas CNNs may just check if features exist on a face (such as a nose, mouth, and eyes), capsule networks determine if they exist in the correct spatial relationships with each other (the nose is above the mouth, and so on), meaning that capsule networks often outperform CNNs on object classification tasks. Using a capsule network requires fewer parameters than CNNs, thus requiring less computing power, and it also can often be more accurate as it takes positionality, not just existence, into stronger consideration. The accuracy of this method for detecting deepfakes was about 92.17%.[63]

9. Dang et al. (2019) utilizes an attention map (a feature map of the section of the image that the model is using the most "attention" to look at) to highlight the regions of the image that most influence the CNN's decision as to whether or not the image is fake. This can be used to further guide the CNN as to which features are most discriminative and significant in determining veracity, making the CNN stronger. This method achieved an AUC of 98.4% on data that was of the same dataset that the model was trained on, and an AUC of 71.2% on data that was of a different dataset.[64] The disparity among results using different datasets implies that training on political deepfakes may not have strong results for verifying pornographic deepfakes. Thus, the training dataset used should encompass the scope of the issue – in this case, should include pornographic deepfakes.

10. Wang and Dantcheva et al. (2020) compare three 3D CNN models used for action recognition and analysis. These methods perform well on deepfakes that use deepfake generation techniques that they were trained on (accuracies ranging from 91% to 95%), but less well on deepfake generation techniques that were not included in the training data.[65]

11. Güera and Delp (2018) offer a detection system based on exploiting three key weaknesses of deepfakes: multiple camera views or differences in lighting causing swapped faces to be inconsistent with the colors, lighting, or angle of the rest of the scene; boundary effects due to a seamed fusion between the swapped face and the rest of the scene; and the frame-by-frame encoding nature of the deepfake development that may cause color and lighting inconsistencies among different frames of the same face, causing a "flickering" effect. Their model obtains a set of features for each frame from the CNN, so the concatenation of the features of multiple frames

[63]Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen, "Use of a Capsule Network to Detect Fake Images and Videos," *ArXiv:1910.12467 [Cs],* October 29, 2019, http://arxiv.org/abs/1910.12467.

[64]Hao Dang et al., "On the Detection of Digital Face Manipulation," *ArXiv:1910.01717 [Cs]*, October 23, 2020, http://arxiv.org/abs/1910.01717.

[65]Yaohui Wang and Antitza Dantcheva, "A Video Is Worth More than 1000 Lies. Comparing 3DCNN Approaches for Detecting Deepfakes," June 9, 2020, https://hal.inria.fr/hal-02862476.

are passed into their detection model to determine the likelihood of the sequence being that of a deepfake.[66]

The strength and breadth of these varied models show a promising future for deepfake detection. However, successful deepfake detection requires not only the initial detection of deepfake videos, but also the continuous detection of distribution of these videos. Thus, once a video has been detected as a deepfake pornography video, online platforms should create a fingerprint for the video that is stored and checked against in the future. For example, if a data storage website such as Google and Microsoft detects a deepfake pornography video being shared, it should add a hash of said video to an internal database. When users share or distribute videos in the future, hashes can be checked against those in the database to ensure that the video is not shared further. Ideally, this database of hashes (which cannot be decrypted back into the videos, so no deepfake pornography videos would be accessible through the database) would be shared among a variety of companies in a data-sharing agreement to impede the spread of deepfake pornography. The feasibility of this catching all deepfake pornography is still out of reach, as videos could be compressed or slightly edited to change the hash in order to evade the database checks, but it would assist in catching deepfake pornography that hasn't been further altered since its last distribution, and future innovations to this technology may assist in catching compressed videos as well.

[66] David Güera and Edward J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," in *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS),* 2018, https://doi.org/10.1109/AVSS.2018.8639163.

4 Potential Legal and Policy Solutions

While the technical detection methods offered in the previous section may engender optimism, we must realize that detection without removal offers little solace to those exploited by deepfake pornography. Detection is an essential initial step, but now we must move on to the steps toward removal.

I offer three paths towards progress: civil liability in the form of tort law and copyright law, personal criminal liability, and platform liability. Civil liability allows for victims to apply tort law violations to the creators of deepfake pornography. Criminalization of deepfake pornography allows for a more codified prohibition of the creation and distribution of deepfake pornography through state or federal laws. Platform liability can exist in the form of user policy restrictions – either created by the platforms themselves or in conjunction with legislative bodies – that disallows distribution of deepfake pornography.

4.1 Civil Liability

Those exploited by deepfake pornographic videos have potential recourse in the form of civil lawsuits, wherein plaintiffs can apply tort law violations to malicious actions that cause distress, invasions of privacy, or harm to one's reputation. California law AB 602, which I discuss in the State Laws Related to Deepfakes section, codifies the ability of victims of nonconsensual deepfake pornography to sue for damages. Outside of California, other civil liability methods are still feasible, though difficulties exist in the form of attribution, jurisdiction issues, financial costs, and social costs. Creators and distributors of deepfakes may take precautionary measures to remain anonymous, such as obfuscating their IP addresses,[67] rendering identification of the person who generated the deepfake infeasible, thus making civil liability impossible. The creator of the deepfake may also be outside of the United States, so civil remedies cannot be used effectively.[68] Even if perpetrators can be identified, civil suits are both expensive and potentially reputationally harmful, especially if the deepfake is embarrassing.[69] Victims may choose not to endure the financial or emotional strain of suing. However, if the creator can be identified, the creator is within a valid jurisdiction, and the victim is willing and able to sue, I offer potential tort violations that can be invoked, including: appropriation of name or likeness, intrusion upon seclusion, public disclosure of private facts, defamation, false light, and intentional infliction of emotional distress.

[67] Citron and Chesney, "Deep Fakes: A Looming Challenge," 1792.
[68] Ibid.
[69] Id., 1792-1793.

Appropriation of Name or Likeness

The appropriation of name or likeness tort applies to situations in which the defendant appropriates another's name or likeness for his or her own use or benefit.[70] While the most common invocation of this rule is in situations of appropriation for the purpose of advertising a business or product, it can also apply to situations where the benefits are not for the purpose of marketing[71] – for instance, the creation of a deepfake video that will be sold in exchange for bitcoins on a deepfake-specific pornographic website[72] – or even in situations where the benefits are not pecuniary.[73] In cases like this, the idea of "benefits" could be understood as notoriety, entertainment, or the use of the video as a form of currency to exchange for other pornographic videos in chatrooms and forums online. If the plaintiff can prove benefits – monetary or otherwise – then this tort would be a viable approach.

Intrusion Upon Seclusion

This tort, commonly used in privacy cases, covers cases in which the defendant "intrudes, physically or otherwise, upon the solitude or seclusion of another or his private affairs or concerns... if the intrusion would be highly offensive to a reasonable person."[74] This may sound applicable at first read as the creation of deepfake videos may be considered to be an intrusion on the plaintiff's private affairs, but this tort is typically only used in the case of physical intrusions on one's affairs, such as opening one's mail, tapping one's telephone wires, forcing entry to one's hotel room, or examining one's private bank account.[75] Since there is no actual physical intrusion to the plaintiff's personal affairs, this tort is not viable for deepfake video cases.

Public Disclosure of Private Facts

Public disclosure of private facts is invoked when the defendant "gives publicity to a matter concerning the private life of another," with the matter being "highly offensive to a reasonable person" and "not of legitimate concern to the public."[76] From this description, one could argue that publishing a video

[70] Restatement (Second) of Torts § 652C (1977).
[71] Restatement (Second) of Torts § 652C cmt. b (1977).
[72] Amrita Khalid, "Deepfake Videos Are a Far, Far Bigger Problem for Women," *Quartz*, accessed December 17, 2020, https://qz.com/1723476/deepfake-videos-feature-mostly-porn-according-to-new-study-from-deeptrace-labs/.
[73] Restatement (Second) of Torts § 652C cmt. b (1977).
[74] Restatement (Second) of Torts § 652B (1977).
[75] Restatement (Second) of Torts § 652B cmt. b. (1977).
[76] Restatement (Second) of Torts § 652D (1977).

online can be considered "giving publicity," and the sexual behaviors of the plaintiff can certainly be considered his or her private life, which is why this tort is commonly used in cases regarding revenge porn.[77] However, it would not be a viable option for a deepfake video case, as the deepfake videos' doctored nature means they are not considered statements of fact. Thus, as this tort provides liability for damages related to publicity given only to true statements of fact,[78] this tort is likely not applicable.

Defamation

The defamation tort can be invoked if the plaintiff can prove that the communication (in this case, the deepfake video) "harm[s] the reputation [of the plaintiff] as to lower him in the estimation of the community or to deter third persons from associating or dealing with him."[79] Thus, in order to invoke this tort, one must prove that the deepfake video harmed her reputation and caused others' estimations of that plaintiff to lower. In the Societal Impacts section of this paper, I show the harms suffered by a plaintiff in deepfake porn cases, including harm to one's career, mental health, and from third parties and/or ex-partners. However, one caveat complicates the usage of defamation in a deepfake case: the truth defense. Truth is a complete defense against defamation,[80] so if a watermark appears on the video stating that the video is doctored, the defendant may be able to utilize this to show truth and negate the possibility of a defamation claim.

False Light

The tort of false light covers the defendant "giv[ing] publicity to a matter concerning another that places the other before the public in a false light," requiring that the false light is "highly offensive to a reasonable person" and was done with either disregard to or knowledge of the "falsity of the publicized matter and the false light in which the other would be placed."[81] In false light cases, the plaintiff can receive damages for the emotional harm suffered due to the falsehood.[82] This may be more applicable than a defamation claim, as a

[77] Caroline Drinnon, "When Fame Takes Away the Right to Privacy in One's Body: Revenge Porn and Tort Remedies for Public Figures," *William & Mary Journal of Race, Gender, and Social Justice* 24, no. 1 (November 2017): 220.

[78] Restatement (Second) of Torts § 652D Special Note (1977).

[79] Restatement (Second) of Torts § 559 (1977).

[80] "Defamation vs. False Light: What Is the Difference?," Findlaw, December 5, 2018, https://www.findlaw.com/injury/torts-and-personal-injuries/defamation-vs–false-light–what-is-the-difference-.html.

[81] Restatement (Second) of Torts § 652E (1977).

[82] "False Light," LII / Legal Information Institute, accessed December 17, 2020, https://www.law.cornell.edu/wex/false_light.

defamation claim may be hindered be the necessity of falsity of the communication. As stated above, if the maker of the video adds a watermark stating that the video is doctored, then the defendant could claim that the video is not false. However, even if the video is not false, it could still place the plaintiff in a false light, making this claim applicable.

Intentional Infliction of Emotional Distress

Intentional infliction of emotional distress (IIED) covers cases in which the defendant partakes in "extreme and outrageous conduct intentionally or recklessly caus[ing] severe emotional distress" to the plaintiff.[83] The emotional distress caused by seeing oneself in a doctored pornographic video, paired with the emotional distress caused by the reactions (both potential and real) of others seeing such a video shows two elements of IIED: "actual suffering of severe and extreme emotional distress," and "actual and proximate causation of the emotional distress by the defendant's outrageous conduct."[84] The lack of consideration of these emotional distresses displays the third element of IIED: "the intention of causing, or reckless disregard of the probability of causing, emotional distress."[85] The final element to consider is the most difficult: "outrageous behavior by the defendant."[86] The difficulty in proving this element of IIED lies in the subjective nature of "outrageous behavior" and the desire to adhere to the First Amendment's allowance of free speech. Two cases, *Texas v. Johnson* (1989) and *Snyder v. Phelps* (2011), display the difficulties in proving the outrageous conduct element. As stated by the *Texas v. Johnson Court*, "government may not prohibit the expression of an idea simply because society finds the idea itself offensive or disagreeable."[87] Thus, merely because society may find that deepfake pornographic videos are offensive or disagreeable does not satisfy the outrageous behavior element. This is further illustrated by the 2011 decision, *Snyder v. Phelps*, in which the Supreme Court disagreed with a jury's verdict that invoked IIED liability for Westboro Baptist picketers at a military funeral. The Supreme Court's decision relied upon its finding that applying the IIED tort "would pose too great a danger that the jury would punish [the defendant] for its views on matters of public concern."[88] Once again, this shows an aversion of the court towards allowing personal or societal views to determine the nature of what may be considered "outrageous conduct", making this tort difficult to invoke.

---

[83] Restatement (Second) of Torts § 46 (1977).

[84] "Intentional Infliction of Emotional Distress," *UpCounsel*, accessed December 17, 2020, https://www.upcounsel.com/lectl-intentional-infliction-of-emotional-distress-tort-law-basics.

[85] Ibid.

[86] Ibid.

[87] Texas v. Johnson, 491 U.S. 397 (1989).

[88] Snyder v. Phelps, 562 U.S. 443 (2011).

Overall Viability of Tort Liability

In summation, I contend that intrusion upon seclusion and public disclosure of private facts are not applicable to deepfake pornography cases; intentional infliction of emotional distress and defamation both have nuances that cause them to be possible to utilize, though potentially difficult to prove; and appropriation of name or likeness and false light are the most applicable means of recourse for those suing for deepfake pornographic exploitation.

Copyright Law

In the discussion of how to mitigate against revenge porn, Citron offers the potential invocation of copyright law. She claims that victims could file notice and takedown requests with content platforms after registering the copyright of their images or videos; once they have been registered, content platforms would then have to take down the pornography promptly or face monetary damages under the Digital Millenium Copyright Act.[89] While Citron argues that this may not be entirely effective for revenge porn as some platforms may ignore requests due to victims' lacking money and/or desire to sue, it seems to be even less effective against deepfake pornography. The "performance" in deepfake pornography is not owned by the victim, but rather by pornographic actors who took part in the original video (or the agency that created it). Even the face transplanted onto the "performance" is usually created from a dataset of publicly available images. Thus, copyright damages would likely not hold for deepfake pornography videos.

One potential way to utilize copyright law to victims' advantage would be if an agreement was created between victims and pornography actors. In an act of solidarity, actors may agree to file the notice and takedown requests for videos in which their performances were used with a victim's face transplanted on.

[89] Citron, "Sexual Privacy," 1935.

4.2 Criminalization of Deepfake Pornography

4.2.1 State Laws Related to Deepfakes

Currently, six states – Virginia, California, Texas, Maryland, Massachusetts, and New York – have passed or proposed laws relating to deepfake videos. Even though 96% of deepfakes online are pornographic, only two of the seven passed or proposed laws are related specifically to deepfake pornography. The percentage of deepfake videos that are pornographic versus the percentage of deepfake laws that are pornography-specific are far from proportional. This disparity exhibits a common thread throughout this paper: deepfake pornography is the primary abuse of deepfake technology, yet due to the demographics of the victims, it is not the most legislated.

This disparity may be due in part to the demographics of those creating deepfake-specific laws versus the demographics of those being exploited by deepfake pornography. As I discuss in the Societal Impacts section of this paper, 100% of the deepfake pornography videos found in the Deeptrace Labs report featured women's faces being added into the video, making women essentially the exclusive victims of this exploitation. As of February 2021, only 30.8% of state legislature seats are held by women, and 26.5% of Congressional seats.[90] The average age of state lawmakers is 56 years old and the average age of Congresspeople is 59 years old.[91] Thus, as the most common American lawmaker is a male from the Baby Boomer generation, legislators may find it difficult to see themselves as victims of deepfake pornography, thus making them prioritize it less than political deepfakes, for example, which would be more likely to target them. The constituency that is most exploited by deepfake pornography has less political representation, which leads to less legal protection.

The disproportionate laws surrounding deepfake pornography may also have to do with media coverage: "Of all worldwide online news coverage of deep fakes in the 65 languages monitored by the GDELT Project, roughly 42% has mentioned its pornographic applications, while 48% has mentioned its potential use in politics."[92] The disproportionately smaller percentage of pornographic-specific media attention may be due in part to the distasteful nature of pornography being mentioned in media outlets, and also due in part to the relatively

[90] "Current Numbers," Rutgers Center for American Women and Politics, June 12, 2015, https://cawp.rutgers.edu/current-numbers.

[91] Karl Kurtz, "Who We Elect: The Demographics of State Legislatures," National Conference of State Legislatures, accessed February 5, 2021, https://www.ncsl.org/research/about-state-legislatures/who-we-elect.aspx.

[92] Kalev Leetaru, "DeepFakes: The Media Talks Politics While The Public Is Interested In Pornography," *Forbes,* March 16, 2019, https://www.forbes.com/sites/kalevleetaru/2019/03/16/deepfakes-the-media-talks-politics-while-the-public-is-interested-in-pornography/.

recent attention surrounding election misinformation. Since media coverage is proportionally less for pornographic deepfakes, it follows that elected officials would pay less attention to deepfake pornography when creating deepfake laws.

State Deepfake Pornography Laws

In March 2019, the General Assembly of Virginia amended § 18.2-386.2 of the Code of Virginia relating to unlawful dissemination or sale of images of another person. This law, used in cases of non-consensual pornography, originally stated: "Any person who, with the intent to coerce, harass, or intimidate, maliciously disseminates or sells any videographic or still image created by any means whatsoever that depicts another person who is totally nude, or in a state of undress so as to expose the genitals, pubic area, buttocks, or female breast, where such person knows or has reason to know that he is not licensed or authorized to disseminate or sell such videographic or still image is guilty of a Class 1 misdemeanor."[93] The amendment to the law included the following: "For purposes of this subsection, 'another person' includes a person whose image was used in creating, adapting, or modifying a videographic or still image with the intent to depict an actual person and who is recognizable as an actual person by the person's face, likeness, or other distinguishing characteristic."[94] Virginia simply added a deepfake-specific provision onto their preexisting nonconsensual pornography law. This is an efficient, effective, and simple way to codify the prohibition of deepfake pornography as law.

In October 2019, California enacted AB 602, which allows a victim of nonconsensual deepfake pornography to sue for damages. The bill states that the depicted individual "has a cause of action against a person who does either of the following: (1) Creates and intentionally discloses sexually explicit material and the person knows or reasonably should have known the depicted individual in that material did not consent to its creation or disclosure. (2) Intentionally discloses sexually explicit material that the person did not create and the person knows the depicted individual in that material did not consent to the creation of the sexually explicit material."[95] In the situations described, the plaintiff may recover "An amount equal to the monetary gain made by the defendant from the creation, development, or disclosure of the sexually explicit material," punitive damages, attorney's fees and costs, "economic and noneconomic damages proximately caused by the disclosure of the sexually explicit material, including damages for emotional distress," "an award of statutory damages for all

[93] "Chapter 515: Virginia's Legislative Information System Bill Tracking," approved March 18, 2019, https://lis.virginia.gov/cgi-bin/legp604.exe?191+ful+CHAP0515&191+ful+CHAP0515.

[94] "Chapter 515: Virginia's Legislative Information System Bill Tracking."

[95] AB 602, "Depiction of Individual Using Digital or Electronic Technology: Sexually Explicit Material," approved October 3, 2019, text from: https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB602.

unauthorized acts involved in the action," and any other available relief.[96]

The deepfake pornography laws in Virginia and California are both undoubtedly significant steps. The path taken by Virginia seems to be the simplest route to the criminalization of deepfake pornography for the majority of states. Currently, 46 states and DC have laws regarding non-consensual pornography, though the states' laws are of varying classifications and punishments. The only states without non-consensual pornography laws as of February 2021 are Massacushetts, Mississippi, South Carolina, and Wyoming.[97] States with pre-existing non-consensual pornography laws can simply add text similar to the Virginia addendum in order to ensure that non-consensual pornography covers the realm of deepfake pornography as well.

State Political Deepfake Laws

More common than statutes related to deepfake pornography are statutes related to political deepfakes. Although non-pornographic political deepfakes only make up at most 4% of the deepfake videos online, two states, Texas and California, currently have laws surrounding political deepfake videos and one state, Maryland, has a pending political deepfake bill in progress as of December 2020. Texas was the first state to prohibit the creation and distribution of deepfakes intended to harm candidates for public office or influence elections, then California also passed a bill making it illegal to "circulate deepfake videos, images, or audio of politicians within 60 days of an election."[98] Maryland has a pending political deepfake bill that may prohibit individuals from "willfully or knowingly influencing or attempting to influence a voter's decision to go to the polls or to cause a vote for a particular candidate by publishing, distributing, or disseminating a deepfake online within 90 days of an election."[99]

Other State Deepfake Laws

Massachusetts and New York have both proposed deepfake laws that are not specific to either pornography or politics. Massachusetts's pending legislation criminalizes the use of deepfakes in conjunction with already "criminal or tortious conduct."[100] Thus, since extortion is illegal in Massachusetts,[101] using a deepfake in order to perform extortion would make the usage of deepfakes illegal as well. New York's pending deepfake law focuses on a different aspect of

[96] Ibid.

[97] "46 States + DC + One Territory NOW Have Revenge Porn Laws — Cyber Civil Rights Initiative," *Cyber Civil Rights Initiative,* accessed February 5, 2021, https://www.cybercivilrights.org/revenge-porn-laws/.

[98] David Ruiz, "Deepfakes Laws and Proposals Flood US," Malwarebytes Labs, January 23, 2020, https://blog.malwarebytes.com/artificial-intelligence/2020/01/deepfakes-laws-and-proposals-flood-us/.

[99] Ibid.

[100] Ibid.

[101] Mass. Gen. Laws ch. 265, § 25. https://malegislature.gov/Laws/GeneralLaws/PartIV/TitleI/Chapter265/Section25.

deepfakes – specifically, the rights to an individual's digital likeness up to forty years after his or her death.[102] This will likely be used in the case of actors and performers whose likeness will be able to be registered to their family members to control whether deepfakes will be allowed in posthumous productions.

4.2.2 Federal Laws Related to Deepfakes

In addition to state-specific statutes, two deepfake-related acts have been passed on a federal level: the National Defense Authorization Act (NDAA) for Fiscal Year 2020 and the Identifying Outputs of Generative Adversarial Networks (IOGAN) Act. Neither of these federal laws are related to deepfake pornography. Rather, the NDAA focuses on the potential national security impacts of deepfakes and establishes a program to award prizes to innovative deepfake detection methods,[103] and the IOGAN Act was created to support research on deepfakes, specifically "technologies for verifying the authenticity of information and detection of manipulated or synthesized content," "watermarking systems for generated media," "social and behavioral research related to manipulated or synthesized content," and "research on public understanding and awareness of manipulated and synthesized content, including research on best practices for educating the public to discern authenticity of digital content."[104]

While neither of these laws specifically mention deepfake pornography and instead focus on political deepfakes or deepfake technology in general, they both support research and funding for deepfake detection systems. Deepfake detection systems would benefit not only the realm of political deepfakes, but also deepfake pornography. These laws, if fulfilled successfully, may lead to stronger deepfake detection tools which can be used by platforms such as Facebook and large pornography aggregators such as Pornhub who have banned deepfake pornography. However, deepfake pornography detection does not assist the issue of deepfake-specific pornography websites, as detection is not the obstacle for those websites. The differences in deepfake pornography enforcement between aggregators and deepfake-specific pornography websites is explained in more depth in the Platform Liability: Pornography Websites section.

[102] Ruiz, "Deepfake Laws and Proposals Flood US."

[103] "Public Law 116-92: National Defense Authorization Act for Fiscal Year 2020," (Stat. 1198; Date: 12/20/2019), text from: Congress.com, accessed February 8, 2021, https://www.congress.gov/bill/116th-congress/senate-bill/1790/text.

[104] "Public Law 116-258: Identifying Outputs of Generative Adversarial Networks Act," (Stat. 1150; Date: 12/23/2020), text from: Congress.com, accessed February 8, 2021, https://www.congress.gov/bill/116th-congress/senate-bill/2904/text.

4.2.3 Impacts to Free Speech

Some may read the above section regarding the criminalization of deepfake pornography and find themselves growing wary of the potential impacts to free speech. The banning and/or censoring of any mode of expression, especially one that may be perceived as "creative," is sure to raise concern. In fact, the California deepfake law AB 602 has already been criticized by the California News Publishers Association and the ACLU for its potential harmful impacts to free speech.[105] We must address these concerns if we are to create an equitable, just action regarding deepfake pornography.

In upholding the First Amendment, we must be clear on what the First Amendment is actually intended to protect. In discussing the caveats of the First Amendment, many have likely heard the hypothetical of yelling "Fire!" in a crowded movie theater, but this is far from the only unprotected form of speech. Protected speech does not include the following categories: obscenity, defamation, fraud, incitement, fighting words, true threats, speech integral to criminal conduct, and child pornography.[106]

Is deepfake pornography protected speech? One could argue that deepfake pornography falls into the "obscene" category, making it unprotected speech. To be labeled as "obscene," it must be judged by modern standards to "appeal to the prurient interest in sex," depict or describe sexual conduct in an offensive way, and lack "serious literary, artistic, political, or scientific value."[107] According to these categories, deepfake pornography could potentially be considered obscene. However, I am apprehensive of applying an outdated, declining categorization to an innovative, novel technological concept. Obscenity prosecutions are in a steady decline and have been since at least the 1990s.[108] Additionally, labeling "deepfake pornography" as obscene may imply that the obscenity stems solely from the pornographic aspect of it. I would like to make it clear that pornography need not be deemed inherently obscene. A danger in banning deepfake pornography may be the implication that pornography itself is a disgusting or obscene practice as it degrades the audience or viewing public, but – though this may be contested – I believe that ethical pornography can and does exist in the form of consensual, legal pornography that respects its workers and their boundaries, celebrates sexual diversity, pays workers, and portrays non-violent, consensual, and non-coercive scenes. Deepfake pornography is not

[105] K.C. Halm et al., "Two New California Laws Tackle Deepfake Videos in Politics and Porn — Davis Wright Tremaine," Davis Wright Tremaine LLP, accessed February 5, 2021, https://www.dwt.com/insights/2019/10/california-deepfakes-law.

[106] Victoria L Killion, "The First Amendment: Categories of Speech," *Congressional Research Service,* January 16, 2019, 2.

[107] Miller v. California, 413 U.S. 15, 24 (1973).

[108] "Obscenity and the First Amendment," Cornell University Law School, accessed February 8, 2021, https://courses2.cit.cornell.edu/sociallaw/student_projects/ObscenityFirstAmendment.htm.

dangerous simply because of the pornographic aspects of it or because of its effects on the viewers, but because of the severe exploitation of the victim's bodily autonomy and invasion of her sexual privacy with no consent whatsoever – for the act itself, for the creation of a video, for the use of her likeness, or for the distribution.

Instead of categorizing deepfake pornography as obscene, I argue that deepfake pornography should follow the precedent set by child pornography in the creation of a new unprotected category. In *New York v. Ferber* (1982), the Supreme Court upheld a New York statute banning the distribution of material depicting sexual performances by children under the age of 16. In the case, the defendant charged with distributing the material argued that the statute violated the First Amendment, but the Court determined that the statute did not violate the First Amendment since child pornography was not protected speech. However, they determined this not by classifying child pornography as obscene, but by creating a new category altogether. This decision hinged on their position that child pornography was distinct from other forms of pornography due to the extreme, harmful "physiological, emotional, and mental" effects to the child. Additionally, they state that the definition of obscenity codified in *Miller v. California*, 413 U.S. 15 is simply not satisfactory in defining the scope of child pornography's atrocity. Finally, they posit that "recognizing and classifying child pornography as a category of material outside the First Amendment's protection is not incompatible with this Court's decisions dealing with what speech is unprotected."[109] We can use this legal precedent to show that similarly, recognizing and classifying deepfake pornography as a category of material outside the First Amendment's protection is not incompatible either.

I also find it important to note that decrying the prohibition of deepfake pornography as a matter of free speech is a distinctly male-focused argument. As discussed in the Societal Impacts section, those whose images are used in deepfake pornography are basically exclusively female, meaning that the creators of the pornography are at least majority male (once again, I would like to note that the Deeptrace Labs study did not include nonbinary genders in the report, and the assumption that the creators of the pornography are majority male is based on a heteronormative concept of the male/female intimate relationship, though I believe that this heteronormative approach likely fits in this scenario). Thus, arguing that the freedom of expression in creating these videos should be prioritized over the digital exploitation of the bodies used in the videos is a marked prioritization of male creators over female victims. In abstract discussions of the impacts of these policies, we should not lose sight of the practical, personal impacts towards women's bodies.

Thus, I argue that impacts to the First Amendment are not applicable in the ban of deepfake pornography, as deepfake pornography should not be

[109] New York v. Ferber, 458 US 747 (1982).

considered protected speech following the precedent set by child pornography. However, for the sake of hypothetical, one could argue that deepfake pornography *should* be protected speech. In that case, being protected simply means that the government cannot infringe upon the speech, not that companies cannot censor or ban the expression. This brings us to our next section concerning online platforms' roles in the censorship of deepfake pornography.

4.3 Platform Liability

4.3.1 Current State of Platform Restrictions

Social Media Websites

Citron and Chesney argue that websites' terms of services agreements are the "single most important documents governing digital speech in today's world, in contrast to prior ages where the first Amendment provided the road map for speech that was permissible in public discourse."[110] To understand the current state of digital speech surrounding deepfakes and deepfake pornography, we must take a look at terms of service agreements on various platforms.

Although deepfake pornography was birthed on Reddit,[111] it was banned soon after. Reddit's user policy now "prohibits the dissemination of images or video depicting any person in a state of nudity or engaged in any act of sexual conduct apparently created or posted without their permission, including depictions that have been faked."[112] While Twitter allows pornographic content to be posted on their platform, it bans any non-consensual nudity, specifically stating that "images or videos that superimpose or otherwise digitally manipulate an individual's face onto another person's nude body" are in violation of Twitter's policies.[113] Facebook has gone further than solely prohibiting deepfake pornography from being shared on their platform; they have banned all manipulated media that are "the product of artificial intelligence or machine learning that merges, replaces or superimposes content onto a video, making it appear to be authentic"[114] – in other words, banning all deepfakes.

Adding deepfake pornography-specific sections to community guidelines on social media sites such as the ones enumerated above is a strong first step for platforms, however the enforcement of these guidelines must be ensured. Under current legislation, as I soon discuss in the Platform Liability Challenges section, bans on deepfake pornography have no means of strict enforcement, whether they are banned in the platform's community guidelines or not.

Pornography Websites

Pornography websites are generally divided into two types: large aggre-

[110] Citron and Chesney, "Deep Fakes: A Looming Challenge," 1817.
[111] Cole, "Deepfakes Were Created As a Way to Own Women's Bodies."
[112] "Do Not Post Involuntary Pornography," Reddit Help, accessed February 5, 2021, https://reddit.zendesk.com/hc/en-us/articles/360043513411-Do-Not-Post-Involuntary-Pornography.
[113] "Non-Consensual Nudity Policy," Twitter Help Center, accessed February 5, 2021, https://help.twitter.com/en/rules-and-policies/intimate-media.
[114] Monika Bickert, "Enforcing Against Manipulated Media," About Facebook, January 6, 2020, https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/.

gators and smaller, usually more specific and niche platforms. MindGeek is the dominant online pornography company, consisting of subsidiaries that include mainstream pornography websites such as PornHub, RedTube, and YouPorn.[115] MindGeek has been criticized as having a monopoly over the digital pornography industry, as it owns three of the top ten pornography sites[116] and in 2018, brought in over 115 million visitors to its websites every day.[117] "Every day, roughly 15 terabytes worth of videos get uploaded to MindGeek's sites, equivalent to roughly half of the content available to watch on Netflix."[118] However, even with the vast amounts of video content added each day, the number of people moderating the content is few, especially compared to that of YouTube, Facebook, and other platforms. The Financial Times reported that the site only had a "couple of dozen" people screening the content uploaded to the site, though MindGeek refuted that claim without providing an alternate figure.[119] How can porn aggregators get away with less moderation (and thus, ostensibly more abusive content) than social media platforms even though they both have a wide following? The taboo of sexual topics has "allowed MindGeek and other distributors to fall under the radar of regulators. While Google-owned YouTube has been dragged in front of politicians for failing to spot and remove copyrighted videos or outright illegal content of people getting harmed, criticism of MindGeek and other tube sites has been muted."[120]

However, even if MindGeek faces less media and political attention than large social media platforms, they still bear more responsibility and regulation than smaller, deepfake-specific pornography websites. "Mr. Tucker, who has worked with online porn companies since the late 1990s and counts MindGeek as a client, says large and well-established porn sites such as PornHub are 'the most responsible [of pornography websites]... they have too much risk not to adhere to laws.'"[121] PornHub, one subsidiary of MindGeek, has enacted corporate policies to create a less abusive online pornography space: prohibiting non-verified users from uploading videos,[122] banning deepfakes, and not returning any results upon searches for "deepfake," "deep fake," or any plural version.[123] This is due in part to the fact that PornHub has much more to lose than smaller porn sites, as they garner large amounts of money and views that make them hold a significant

[115]Patricia Nilsson, "MindGeek: The Secretive Owner of Pornhub and RedTube," *Financial Times,* December 17, 2020, https://www.ft.com/content/b50dc0a4-54a3-4ef6-88e0-3187511a67a2.
[116]Jess Joho, "The Best Alternatives to Pornhub and Xvideos," *Mashable*, December 15, 2020, https://mashable.com/article/pornhub-alternatives-free-porn-paid-porn/.
[117]Nilsson, "MindGeek: The Secretive Owner of Pornhub and RedTube."
[118]Ibid.
[119]Ibid.
[120]Ibid.
[121]Ibid.
[122]Nicholas Kristof, "Opinion: An Uplifting Update, on the Terrible World of Pornhub," *The New York Times,* December 10, 2020, sec. Opinion, https://www.nytimes.com/2020/12/09/opinion/pornhub-news-child-abuse.html.
[123]Pornhub. Accessed Feb 5, 2021.

place in the public eye. PornHub's willingness to bow to public pressure was made obvious in December 2020 when Nicholas Kristof published a piece in the New York Times exposing the amount of revenge porn and child pornography present in non-verified videos.[124] The decision to ban non-verified users' videos occurred only a few days after the report was posted, an obvious response to the public outcry stemming from the article.

Compared to mainstream pornography websites, deepfake-specific pornography websites, the addresses of which I will not list in this paper, have a much smaller following yet house the vast majority of deepfake pornography online; according to the Deeptrace Labs Report, 94% of deepfake pornography found was hosted on dedicated deepfake pornography sites, compared to only 6% being found on mainstream pornography sites.[125] These websites are more difficult to regulate because they have less to lose than aggregators – fewer viewers, less money, and so on. Also, for obvious reasons, they are not attempting to cleanse their sites of deepfake videos as PornHub is at least ostensibly attempting to do. These websites will be the most difficult to hold liable, but I discuss possibilities in the Next Steps to Augment Platform Liability section.

4.3.2 Platform Liability Challenges

Section 230

Before diving into better ways to enforce platform liability, I begin by explaining its largest obstacle: Section 230 of the Communications Decency Act, a 1996 law that discusses regulation and liability issues of the internet.

Section 230 begins by summarizing Congress's findings on the current state of the internet, positing that interactive computer services offer users control over the information they receive and act as a forum for "a true diversity of political discourse," "opportunities for cultural development," as well as "myriad avenues for intellectual activity." The findings reflect a late-1990's technological optimism regarding the internet, singing its praises by claiming that "the internet and other interactive computer services have flourished, to the benefit of all Americans, with a minimum of government regulation."[126] Now that we have twenty-five years of perspective, this finding, stated as fact in a Congressional Act, should be called into question; how do we define "flourished"? If flourishing is equated to financial success, then yes, the internet has grown to be wildly financially successful – but are there other metrics of success? The

[124] Nicholas Kristof, "Opinion: The Children of Pornhub," The New York Times, December 4, 2020, sec. Opinion, https://www.nytimes.com/2020/12/04/opinion/sunday/pornhub-rape-trafficking.html.
[125] Ajder et al., "The State of Deepfakes," 6.
[126] 47 U.S. Code § 230

internet has not flourished at expunging itself of privacy violations, including violations of users' sexual privacy. The internet has not flourished at cleansing itself of hate speech, cyberbullying, and incitements of violence. The internet has by no means flourished at creating a just, non-abusive, non-exploitative, non-racist, non-sexist environment. When Section 230 states that the internet has flourished to the "benefit of all Americans," we must understand that at this point, "all Americans" does not really mean all Americans – only those who have not and could not potentially be abused online.

Section 230 goes on to detail the United States' policies regarding the internet, stating that the United States will continue to promote the growth of the internet and "preserve the vibrant and competitive free market that presently exists for the Internet and other interactive computer services, unfettered by Federal or State regulation."[127] It then creates a shield for platforms that has been widely contested in the years following, stating: "No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider."[128] This declares that providers of interactive computer services, such as social media platforms or pornography websites, are not liable for any inappropriate or illegal content posted by a user.

Section 230 has been an extremely controversial statute. I will soon explain the dangers of the section, but it would be unrepresentative to leave out a discussion of the benefits of the section as well. The initial purpose was actually to create a safer digital environment by encouraging internet moderation. Before Section 230, websites that displayed offensive or illegal content with no moderation could not be sued for that offensive content, as they were only distributing the content and had no expectations of knowing the obscenity levels of everything they published. However, when websites began to moderate some offensive content, they were no longer simply distributors, but were in control of the content they were distributing. The internet service that provided *some* moderation was liable for that content, whereas the internet services that provided no moderation was not.[129] This caused an obvious issue as it disincentivized moderation whatsoever, so Congress wanted to create a law to encourage moderation, stating that just because websites moderate the content on their pages does not mean they are liable for that content. This has given websites broad discretion to moderate content that is illegal and even content that is not necessarily illegal, such as Facebook banning all manipulated media – also known as deepfakes – even though they are currently legal.

[127] Ibid.
[128] Ibid.
[129] Adi Robertson, "Why Section 230 Exists and How People Are Still Getting It Wrong," *The Verge*, June 21, 2019, https://www.theverge.com/2019/6/21/18700605/section-230-internet-law-twenty-six-words-that-created-the-internet-jeff-kosseff-interview.

However, the unintended impact that Section 230 has on the inability to hold platforms liable for their content is dangerous at best. Citron and Wittes analogize that any business that knowingly allows for sexual abuses to happen in their physical space would not be able to easily avoid liability, yet simply because certain businesses operate in a digital environment means that they are free from liability.[130] The shielding of liability for actions solely because they exist online (where, as we know, a plethora of sexual abuses do exist) is dangerous and not aligned with the original intention of the Section. Citron and Wittes summarize the dangers of the current interpretation of Section 230 as follows: "[Section 230] has produced an immunity from liability that is far more sweeping than anything the law's words, context, and history support. Platforms have been protected from liability even though they republished content knowing it might violate the law, encouraged users to post illegal content, changed their design and policies for the purpose of enabling illegal activity, or sold dangerous products."[131]

While this free, unregulated nature of the internet may have been helpful in its fetal stages, it is far from necessary in the full-fledged internet environment in existence today. Citron and Wittes summarize this as follows: "If a broad reading of the safe harbor embodied sound policy in the past, it does not in the present—an era in which child (and adult) predation on the internet is rampant, cybermobs terrorize people for speaking their minds, and designated foreign terrorist groups use online services to organize and promote violent activities."[132] The internet no longer needs to be hand-held.

Misinterpretation of Future Platform Liability Statutes

Section 230 of the Communications Decency Act also displays another challenge to platform immunity: misinterpretation of privacy and content laws. The goal of Section 230 was to incentivize self-regulation by platforms in order to decrease the amount and accessibility of offensive materials online. However, the current interpretation of Section 230 greatly undermines this original congressional goal. Thus, if new legislation is written, it is imperative to write with specific language so as to not allow broad and contradictory interpretations. This is a major challenge to the creation of new statutes, as it may be impossible to predict the future usage of technology-specific laws once novel technologies have been introduced.

Impacts to Free Speech

The First Amendment impacts of criminalization of deepfake pornography

[130]Danielle Keats Citron and Benjamin Wittes, "The Internet Will Not Break: Denying Bad Samaritans § 230 Immunity," *Fordham Law Review* 86 (2017).
[131]Id., 408-409.
[132]Id., 411.

is discussed in the Criminalization of Deepfake Pornography section, but there is another aspect of censorship that is important to discuss: that by private companies as opposed to the government. In our solutions below, some consist of private companies deciding what to moderate and what to censor from their platform. Some may read this and argue that this is a violation of the users' freedom of speech, so I am taking this space to clarify that the First Amendment protects against *government* infringement on protected speech, not private companies or private individuals' decisions to infringe on any speech, protected or not. In fact, it would be a First Amendment issue to not allow a platform to censor its users (such as Twitter deciding to ban Donald Trumps' account in January 2021) because a social media company has the right to exclude speakers that it chooses not to make available on its platform. Though some could argue that there should be limits to this power, particularly where platforms have monopoly power, those limits almost certainly do not apply to the ability to censor deepfake pornography. Thus, even in the hypothetical situation that deepfake pornography is protected speech, that does not mean that PornHub necessarily has to host it, nor do other companies have to provide technical support to websites that do host it.

4.3.3 Next Steps to Augment Platform Liability

Section 230 Updates

Citron and Wittes give two paths for updating Section 230: an interpretive shift and a legislative shift. The current interpretations for Section 230 are based on state and lower federal court decisions, which have interpreted Section 230 to be construed broadly, whereas the Supreme Court has declined to interpret the meaning of Section 230.[133] However, this over-broad interpretation supported by the state and lower courts is dangerous. Citron and Wittes argue that the interpretive solution to this challenge is to interpret Section 230 in "a manner more consistent with its text, context, and history" by limiting its application to true Good Samaritans, defined in their paper as "providers or users engaged in good faith efforts to restrict illegal activity."[134] Finally, they argue that "[t]reating abusive website operators and Good Samaritans alike devalues the efforts of the latter and may result in less of the very kind of blocking that the CDA in general, and § 230 in particular, sought to promote."[135] In the distinction between abusive website operators and Good Samaritans, one could easily argue that deepfake-specific pornographic websites are in the former group. Thus, the path to update Section 230 via interpretation is for lower courts to begin to interpret Section 230 in this way (as it was originally intended)

[133]Id., 407.
[134]Id., 415-416.
[135]Ibid.

or for the Supreme Court to hear a case regarding the Section and create a more firm interpretation.

Aside from interpretation, Section 230 could also be updated via legislation. One such example is a proposal by the National Association of Attorneys General to amend Section 230 to exempt state criminal laws.[136] If this becomes adopted, then the state laws I discuss in the Criminalization of Deepfake Pornography section would allow for platforms to be liable for those crimes as well. Another alternative would be an amendment to the Section that eliminates immunity for the worst actors online, such as "sites that encourage destructive online abuse or that know they are principally used for that purpose."[137] Citron proposes the following amendment:

> "Nothing in Section 230 shall be construed to limit or expand the application of civil or criminal liability for any website or other content host that purposefully encourages cyber stalking[,] . . . nonconsensual pornography," sex trafficking, child sexual exploitation, or that has knowledge that it principally hosts such material."[138]

She also proposes a broader clarification amendment:

> "No provider or user of an interactive computer service that takes reasonable steps to prevent or address unlawful uses of its services shall be treated as the publisher or speaker of any information provided by another information content provider in any action arising out of the publication of content provided by that information content provider."[139]

She and Wittes explain the benefit of her proposal: "With this revision, platforms would enjoy immunity from liability if they could show that their response to unlawful uses of their services was reasonable. Such a determination would take into account differences among online entities. ISPs and social networks with millions of postings a day cannot plausibly respond to complaints of abuse immediately, let alone within a day or two."[140] Thus, either of these two amendment proposals would allow for a more accurate, specific interpretation of Section 230 that would mitigate the challenges created by its overbroad interpretation. This would eliminate the platform immunity that currently exists, making social media websites and pornography websites accountable for

[136] Id., 418.
[137] Id., 419.
[138] Ibid.
[139] Ibid.
[140] Ibid.

the content they publish on their websites, incentivizing them to quickly delete and ban deepfake pornography.

Platform Agreements

In addition to changing and re-interpreting Section 230, there are platform-specific actions that can be taken to undermine the proliferation of deepfake pornography online. Companies can work on a voluntary basis with the government to suppress exploitative and abusive websites. Some examples of partnerships could consist of:

1. Internet service providers such as Verizon and Comcast can decide not to host websites that publish deepfake pornography, thus resolving to a DNS error when end users attempt to access the website.

2. App stores such as the Google Play Store and the iOS App Store could agree to not approve applications that create deepfake videos unless the application is built to detect and disallow the creation of pornographic videos.

3. Social media websites such as Twitter and Reddit could disallow links leading to known deepfake pornography websites (and work with content moderators to keep the list of websites updated).

4. Search engines such as Google and Bing could decide to not index known deepfake pornography websites, thus leading to them not showing up in searches.

5. Cloud hosting services such as Amazon Web Services and Microsoft Azure could enact a policy that disallows their cloud hosting services to be used on websites that host deepfake pornography in a similar manner to Amazon Web Services' decision to halt the hosting of Parler, citing Parler's "violent content" violating Amazon Web Service's Terms of Service.[141] Their Terms of Service could extend not only to prohibiting content that incites violence, but also prohibiting deepfake pornography as well.

6. Browsers such as Google Chrome, Firefox, and Safari could agree to block users from entering known deepfake pornography websites.

7. Any pornography website such as PornHub could require proof of consent before approving a video to be posted. This would assist in banning not only deepfake pornography, but other revenge pornography as well.

[141] Annie Palmer, "Amazon Drops Parler from Its Web Hosting Service, Citing Violent Posts," *CNBC*, January 9, 2021, https://www.cnbc.com/2021/01/09/amazon-drops-parler-from-its-web-hosting-service.html.

8. Payment-based companies such as PayPal, Mastercard, Visa, and American Express could refuse to provide payment services for websites that host deepfake pornography (PayPal has already suspended cooperation with PornHub,[142] with Mastercard and Visa reviewing their ties with the streaming service).[143]

9. Companies with data storage services such as Google and Microsoft can agree to the deepfake pornography fingerprinting technical mitigation discussed in the Technical Detections section.

These are only nine examples of a wide range of potential voluntary platform agreements. Since these are voluntary, they would not be required legally, but could be encouraged by lobbyists, activists, and the general public. We have seen how public pressure towards PornHub has led them to enact stricter policies on the content allowed on their website. Directing energy towards large pornography aggregators is the obvious first step in purging the internet of deepfake pornography, but this should be extended to companies that indirectly host and enable deepfake pornography as well, including internet service providers, cloud hosting services, payment services, search engines, app stores, social media websites, browsers, and many more technical providers. Even with these agreements, deepfake pornography would likely still exist online – it may move further onto the dark web, for example – but would be less easily accessible, and would allow for actions to take it down once detected. This is a strong step in impeding the growth and distribution of deepfake pornography in order to form a less exploitative and abusive internet environment.

[142] Kristof, "The Children of Pornhub."
[143] Kristof, "An Uplifting Update."

Conclusion

Deepfake models utilize deep learning techniques such as neural networks, encoder-decoder pairs, and Generative Adversarial Networks in order to create realistic doctored videos. While not all usages of this technology are harmless, the most widespread form of deepfake videos on the internet are pornographic, and deepfake pornography on the internet usually features everyday women whom the creator knows in real life. Deepfake pornography serves as a means of digital sexual exploitation and virtual sexual abuse, as many victims suffer through the loss of bodily autonomy and self-identity after viewing themselves performing sexual acts which they neither performed nor consented to. It can cause victims to lose trust in future partners and can instill a fear of future children, partners, or employers seeing the video and adjusting their view of the victim's identity or sexuality. While the roots and consequences of deepfake pornography parallel those of a variety of other forms of sexual abuse, deepfake pornography is specifically digital. Because of this, there may be technical and legal ways to mitigate these abuses.

In order to mitigate their exploitative and abusive harms, deepfake pornography should be considered unprotected speech and thus banned from the internet. In order to do so, we must first solve the problem of deepfake detection. There are a variety of methods, eleven of which are explained in this paper, that offer distinct approaches to this problem. As deepfake creation methods are constantly evolving and improving, detection methods must follow suit; an emphasis should be placed on the progress of detection methods as opposed to creation methods in order to perform socially-conscious research. Once a video has been detected, there are a variety of options for recourse the victim: victims can pursue civil liability, criminal liability in some states, and potentially work with platforms to take the content down. However, the current status of these methods are not strong enough. Criminalization of deepfake pornography should be in effect in all states or at a federal level. Platforms should be held accountable, either through updates to Section 230 or through public pressure to adopt the agreements listed in the Next Steps to Augment Platform Liability section to hinder the spread and distribution of deepfake pornography post-detection. Evidence of public pressure creating a more just digital environment can be found in cases like Pornhub's newest restrictions, so users of online platforms should call platforms to action to agree to these steps. Deepfake pornography is a gendered issue that exacerbates the power dynamics of abusive relationships and harrassment, so legal and technical mitigation methods must be put in place to halt the distribution of deepfake pornography in order to create a more just, equitable, and non-abusive digital landscape.

References

Afchar, Darius, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. "MesoNet: A Compact Facial Video Forgery Detection Network." In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, 1–7, 2018. https://doi.org/10.1109/WIFS.2018.8630761.

Agarwal, Shruti, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. "Protecting World Leaders Against Deep Fakes," 2019.

Ajder, Henry, Giorgio Patrini, Francesco Cavalli, and Laurence Cullen. "The State of Deepfakes: Landscape, Threats, and Impact." *Deeptrace,* September 2019.

Bickert, Monika. "Enforcing Against Manipulated Media." About Facebook, January 6, 2020. https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/.

"Chapter 515: Virginia's Legislative Information System Bill Tracking." Approved March 18, 2019. https://lis.virginia.gov/cgi-bin/legp604.exe?191+ful+CHAP0515&191+ful+CHAP0515.

Citron, Danielle Keats, and Benjamin Wittes. "The Internet Will Not Break: Denying Bad Samaritans § 230 Immunity." *Fordham Law Review* 86 (2017).

Citron, Danielle Keats. "Sexual Privacy." *Yale Law Journal* 128, no. 7 (May 1, 2019). https://digitalcommons.law.yale.edu/ylj/vol128/iss7/2.

Citron, Danielle, and Robert Chesney. "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security." *California Law Review* 107, no. 6 (December 1, 2019): 1753.

Cole, Samantha. "Deepfakes Were Created As a Way to Own Women's Bodies—We Can't Forget That." *Vice,* June 18, 2018. https://www.vice.com/en/article/nekqmd/deepfake-porn-origins-sexism-reddit-v25n2.

Cornell University Law School. "Obscenity and the First Amendment." Accessed February 8, 2021. https://courses2.cit.cornell.edu/sociallaw/student_projects/ObscenityFirstAmendment.htm.

Cyber Civil Rights Initiative. "46 States + DC + One Territory NOW Have Revenge Porn Laws — Cyber Civil Rights Initiative." Accessed February 5, 2021. https://www.cybercivilrights.org/revenge-porn-laws/.

Dang, Hao, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil Jain. "On the Detection of Digital Face Manipulation." *ArXiv:1910.01717 [Cs]*, October 23, 2020. http://arxiv.org/abs/1910.01717.

DeepAI. "Convolutional Neural Network," May 17, 2019. https://deepai.org/machine-learning-glossary-and-terms/convolutional-neural-network.

"Deepfake Tech Drastically Improves CGI Princess Leia in Rogue One — GeekTyrant." Accessed March 22, 2021. https://geektyrant.com/news/deepfake-tech-drastically-improves-cgi-princess-leia-in-rouge-one.

Drinnon, Caroline. "When Fame Takes Away the Right to Privacy in One's Body: Revenge Porn and Tort Remedies for Public Figures." *William & Mary Journal of Race, Gender, and Social Justice* 24, no. 1 (November 2017): 220.

"Dynamic Score Combination: A Supervised and Unsupervised Score Combination Method - Scientific Figure." ResearchGate. https://www.researchgate.net/figure/An-example-of-a-ROC-curve-its-AUC-and-its-EER_fig1_225180361.

Findlaw. "Defamation vs. False Light: What Is the Difference?," December 5, 2018. https://www.findlaw.com/injury/torts-and-personal-injuries/defamation-vs--false-light--what-is-the-difference-.html.

GitHub. "Deepfakes/Faceswap/USAGE.Md." Accessed March 22, 2021. https://github.com/deepfakes/faceswap.

Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative Adversarial Networks." *ArXiv:1406.2661 [Cs, Stat]*, June 10, 2014. http://arxiv.org/abs/1406.2661.

Google Developers. "Classification: ROC Curve and AUC — Machine Learning Crash Course." Accessed March 22, 2021. https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc.

Güera, David, and Edward J. Delp. "Deepfake Video Detection Using Recurrent Neural Networks." In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2018. https://doi.org/10.1109/AVSS.2018.8639163.

Halm, K.C., Ambika Kumar Doran, Jonathan Segal, and Caesar Kalinowski IV. "Two New California Laws Tackle Deepfake Videos in Politics and Porn." Davis Wright Tremaine LLP. Accessed February 5, 2021. https://www.dwt.com/insights/2019/10/california-deepfakes-law.

Harwell, Drew. "Fake-Porn Videos Are Being Weaponized to Harass and

Humiliate Women: 'Everybody Is a Potential Target.'" *The Washington Post,* December 30, 2018. https://www.washingtonpost.com/technology/2018/12/30/fake-porn-videos-are-being-weaponized-harass-humiliate-women-everybody-is-potential-target/.

Havron, Sam, Diana Freed, Nicola Dell, Rahul Chatterjee, Damon McCoy, and Thomas Ristenpart. "Clinical Computer Security for Victims of Intimate Partner Violence."

Johnson, Robert and Adam Cureton. "Kant's Moral Philosophy." The Stanford Encyclopedia of Philosophy (Spring 2021 Edition). Edited by Edward N. Zalta. https://plato.stanford.edu/archives/spr2021/entries/kant-moral/.

Joho, Jess. "The Best Alternatives to Pornhub and Xvideos." *Mashable*, December 15, 2020. https://mashable.com/article/pornhub-alternatives-free-porn-paid-porn/.

Jung, Tackhyun, Sangwon Kim, and Keecheon Kim. "DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern." *IEEE* Access 8 (2020). https://doi.org/10.1109/ACCESS.2020.2988660.

Khalid, Amrita. "Deepfake Videos Are a Far, Far Bigger Problem for Women." *Quartz*. Accessed December 17, 2020. https://qz.com/1723476/deepfake-videos-feature-mostly-porn-according-to-new-study-from-deeptrace-labs/.

Killion, Victoria L. "The First Amendment: Categories of Speech." *Congressional Research Service,* January 16, 2019, 2.

Korshunov, Pavel, and Sebastien Marcel. "DeepFakes: A New Threat to Face Recognition? Assessment and Detection." *ArXiv:1812.08685 [Cs]*, December 20, 2018. http://arxiv.org/abs/1812.08685.

Kristof, Nicholas. "Opinion — An Uplifting Update, on the Terrible World of Pornhub." *The New York Times,* December 10, 2020, sec. Opinion. https://www.nytimes.com/2020/12/09/opinion/pornhub-news-child-abuse.html.

Kristof, Nicholas. "Opinion — The Children of Pornhub." *The New York Times*, December 4, 2020, sec. Opinion. https://www.nytimes.com/2020/12/04/opinion/sunday/pornhub-rape-trafficking.html.

Krose, Ben, and Patrick van der Smagt. "An Introduction to Neural Networks," November 1996. http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=1DC5C0944BD8C52584A8C2310AE64181?doi=10.1.1.18.493&rep=rep1&type=pdf.

Kurtz, Karl. "Who We Elect: The Demographics of State Legislatures." National Conference of State Legislatures. Accessed February 5, 2021. https://www.ncsl.org/research/about-state-legislatures/who-we-elect.aspx.

Leetaru, Kalev. "DeepFakes: The Media Talks Politics While The Public Is Interested In Pornography." *Forbes*, March 16, 2019. https://www.forbes.com/sites/kalevleetaru/2019/03/16/deepfakes-the-media-talks-politics-while-the-public-is-interested-in-pornography/.

LII / Legal Information Institute. "False Light." Accessed December 17, 2020. https://www.law.cornell.edu/wex/false_light.

Lomas, Natasha. "Deepfake Video App Reface Is Just Getting Started on Shapeshifting Selfie Culture — TechCrunch." Accessed March 22, 2021. https://techcrunch.com/2020/08/17/deepfake-video-app-reface-is-just-getting-started-on-shapeshifting-selfie-culture/.

Matern, Falko, Christian Riess, and Marc Stamminger. "Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations." In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, 83–92, 2019. https://doi.org/10.1109/WACVW.2019.00020.

Nguyen, Huy H., Junichi Yamagishi, and Isao Echizen. "Use of a Capsule Network to Detect Fake Images and Videos." *ArXiv:1910.12467 [Cs]*, October 29, 2019. http://arxiv.org/abs/1910.12467.

Nguyen, Thanh Thi, Cuong M. Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, and Saeid Nahavandi. "Deep Learning for Deepfakes Creation and Detection: A Survey." *ArXiv:1909.11573 [Cs, Eess]*, July 28, 2020. http://arxiv.org/abs/1909.11573.

Nilsson, Patricia. "MindGeek: The Secretive Owner of Pornhub and RedTube." *Financial Times,* December 17, 2020. https://www.ft.com/content/b50dc0a4-54a3-4ef6-88e0-3187511a67a2.

Nulph, Dr Robert G. "Edit Suite: Once Upon a Time: The History of Videotape Editing." *Videomaker (blog)*, July 1, 1997. https://www.videomaker.com/article/2896-edit-suite-once-upon-a-time-the-history-of-videotape-editing/.

Palmer, Annie. "Amazon Drops Parler from Its Web Hosting Service, Citing Violent Posts." *CNBC*, January 9, 2021. https://www.cnbc.com/2021/01/09/amazon-drops-parler-from-its-web-hosting-service.html.

"Public Law 116-258: Identifying Outputs of Generative Adversarial Networks Act." (Stat. 1150; Date: 12/23/2020). Text from: Congress.com. Accessed

February 8, 2021. https://www.congress.gov/bill/116th-congress/senate-bill/2904/text.

"Public Law 116-92: National Defense Authorization Act for Fiscal Year 2020." (Stat. 1198; Date: 12/20/2019). Text from: Congress.com. Accessed February 8, 2021. https://www.congress.gov/bill/116th-congress/senate-bill/1790/text.

Reddit Help. "Do Not Post Involuntary Pornography." Accessed February 5, 2021. https://reddit.zendesk.com/hc/en-us/articles/360043513411-Do-Not-Post-Involuntary-Pornography.

Robertson, Adi. "Why Section 230 Exists and How People Are Still Getting It Wrong." *The Verge*, June 21, 2019. https://www.theverge.com/2019/6/21/18700605/section-230-internet-law-twenty-six-words-that-created-the-internet-jeff-kosseff-interview.

Ruiz, David. "Deepfakes Laws and Proposals Flood US." Malwarebytes Labs, January 23, 2020. https://blog.malwarebytes.com/artificial-intelligence/2020/01/deepfakes-laws-and-proposals-flood-us/.

Rutgers Center for American Women and Politics. "Current Numbers," June 12, 2015. https://cawp.rutgers.edu/current-numbers.

Scott, Daniella. "Deepfake Porn Nearly Ruined My Life." *Elle,* June 2, 2020. https://www.elle.com/uk/life-and-culture/a30748079/deepfake-porn/.

Singer, P.W., and Allan Friedman. *Cybersecurity and Cyberwar: What Everyone Needs to Know.* Oxford University Press, 2014.

Stuart Mill, John. "The Subjection of Women." In *Feminism: The Essential Historical Writings.* Edited by Miriam Schneir (New York: Vintage Books, 1972).

Suwajanakorn, Supasorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. "Synthesizing Obama: Learning Lip Sync from Audio." *ACM Transactions on Graphics* 36, no. 4 (July 20, 2017): 1–13. https://doi.org/10.1145/3072959.3073640.

Tolosana, Ruben, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. "Deepfakes and beyond: A Survey of Face Manipulation and Fake Detection — Elsevier Enhanced Reader," December 2020. https://doi.org/10.1016/j.inffus.2020.06.014.

Tolosana, Ruben, Sergio Romero-Tapiador, Julian Fierrez, and Ruben Vera-Rodriguez. "DeepFakes Evolution: Analysis of Facial Regions and Fake

Detection Performance." *ArXiv:2004.07532 [Cs]*, July 2, 2020. http://arxiv.org/abs/2004.07532.

Twitter Help Center. "Non-Consensual Nudity Policy." Accessed February 5, 2021. https://help.twitter.com/en/rules-and-policies/intimate-media.

Twitter Help Center. "Sensitive Media Policy." Accessed February 5, 2021. https://help.twitter.com/en/rules-and-policies/media-policy.

UpCounsel. "Intentional Infliction of Emotional Distress." Accessed December 17, 2020. https://www.upcounsel.com/lectl-intentional-infliction-of-emotional-distress-tort-law-basics.

Vincent, James. "Deepfake Bots on Telegram Make the Work of Creating Fake Nudes Dangerously Easy." *The Verge,* October 20, 2020. https://www.theverge.com/2020/10/20/21519322/deepfake-fake-nudes-telegram-bot-deepnude-sensity-report.

Vincent, James. "I Learned to Make a Lip-Syncing Deepfake in Just a Few Hours (and You Can, Too)." *The Verge,* September 9, 2020. https://www.theverge.com/21428653/lip-sync-ai-deepfake-wav2lip-code-how-to.

Vox. *The Most Urgent Threat of Deepfakes Isn't Politics*, 2020. https://www.youtube.com/watch?v=hHHCrf2-x6w&feature=emb_logo.

Wang, Yaohui, and Antitza Dantcheva. "A Video Is Worth More than 1000 Lies. Comparing 3DCNN Approaches for Detecting Deepfakes," June 9, 2020. https://hal.inria.fr/hal-02862476.

Warren, Samuel D., and Louis D. Brandeis. "The Right to Privacy." *Harvard Law Review* 4, no. 5 (1890). https://doi.org/10.2307/1321160.

Woodford, Chris. "How Neural Networks Work - A Simple Introduction." Explain that Stuff, June 17, 2020. http://www.explainthatstuff.com/introduction-to-neural-networks.html.

Yang, Xin, Yuezun Li, and Siwei Lyu. "Exposing Deep Fakes Using Inconsistent Head Poses." In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8261–65, 2019. https://doi.org/10.1109/ICASSP.2019.8683164.

Zhou, Peng, Xintong Han, Vlad I. Morariu, and Larry S. Davis. "Two-Stream Neural Networks for Tampered Face Detection." In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017. https://doi.org/10.1109/CVPRW.2017.229.