It Is Not My Responsibility: Failures in Preventing Malicious Deepfakes

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Wendy Zheng

Spring 2024

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor

Caitlin D. Wylie, Department of Engineering and Society

Introduction

In January 2024, images of Taylor Swift were used to create a pornographic deepfake, gaining over 45 million views and 24,000 reposts on X in about 17 hours before it was taken down (Weatherbed, 2024). Despite the lack of Swift's consent to create the content and the damage it inflicted on her reputation, "there is no direct deepfake federal statute," leaving Swift and other victims of deepfakes to fend for themselves (Contreras, 2024). Deepfakes are fake visual and / or audio content generated by AI to portray the targeted person performing some action they never did. Obviously, they have many dangerous applications, including pornography, misinformation, and fraud. Even so, there are few measures put in place to protect the people from the harmful implications of deepfakes, allowing for the development of malicious deepfakes.

Before taking action against deepfakes, we first need to understand what led to the current prevalence of malicious deepfakes. Bruno Latour's (1996) actor-network theory (ANT) provides a framework to explain a phenomenon through observing the associations between it and other actors, both human and nonhuman ones (pp. 375 - 376). By uncovering the different actors and their relationship to deepfakes, especially malicious ones, we can discover the causes that allowed these malicious deepfakes to develop. I will conduct a document analysis to reveal each actor's stance on deepfakes and the effects of their actions. Although there are many actors in the network of malicious deepfakes, the failures of the U.S. government, software developers, and deepfake creators to take responsibility led to the increasing misuse of deepfakes for malicious activities.

The Government

Due to current law, the government has done little to successfully protect people from deepfakes. Many legal protections can be used to protect the creators of deepfakes, and amending legislation to regulate deepfakes is a difficult task due to fear of violating these protections. Despite some attempts, the government fails to target the root problem of criminalizing illegal deepfakes, rendering their efforts ineffective.

Under strict interpretation of the First Amendment, malicious deepfakes are not included in the freedoms of speech and expression. The amendment protects any form of expression against the government and forbids restrictions by content except under certain extreme circumstances (Stone & Volokh, n.d.). Based on the interpretation from two American Law Professors who specialize on the First Amendment, those circumstances include defamation and obscenity. In regards to deepfakes, the intent of the creator determines if their creation is protected, so the government should supposedly be able to regulate deepfakes that are intentionally malicious. Thus, the amendment prohibits malicious deepfakes by not providing protection for them, making it an actor in the network

However, the Supreme Court is adopting a broader interpretation of the amendment, meaning stronger protection of rights granted by the amendment and stricter scrutiny of laws restricting those rights. This is evident in the many cases that created loopholes for malicious deepfake creators to slip through the legal system. For example, two landmark cases in libel defined how First Amendment protections apply to defamatory statements. *Gertz v. Robert Welch, Inc.* (1974) concluded that publishers of defamatory information cannot be held liable without fault / intent and proof of damage to a private figure's reputation. *New York Times Co. v. Sullivan* (1964) imposed stricter guidelines for when the target is a public figure, requiring evidence of “actual malice,” which means knowingly spreading a false statement. For defamatory deepfakes, regardless if they are targeting a public or private figure, the creator can post them anonymously, making it difficult to find concrete evidence and take legal action against them. The decisions of these cases reflect the Supreme Court’s strong support for the protection of free speech and expression, but the application of precedent cases in modern times can lead to injustice for the victims.

As technology advances, definitions and interpretations of laws need to be updated to ensure that the current system creates a fair and safe environment for its people. With the Supreme Court’s interpretation, people can exploit the situation for personal gains, such as creating malicious deepfakes without fear of legal consequences. In a similar situation, interpretations of the Fourth Amendment with regards to search and seizure changed to acknowledge technological advances: the amendment used to allow the police to search everything they can find on an arrestee, but with the invention of the smartphone, the Supreme Court ruled that the police needed a warrant to search an arrestee’s phone (Blitz, 2020, p. 248). Blitz, a professor of law who studies the implications of new technologies on constitutional law, specifically the First and Fourth amendments, points out that redefining foundational amendments has been done before and suggests the same be done with the First Amendment. By recognizing deepfakes as a technologically advanced form of expression, courts are likely to give more just decisions for the targeted people.

Outside of the legal system, the government has introduced bills against the harms of deepfakes, but all except one failed to become a law and are not enough to protect the victims. The National Defense Authorization Act (2020) is the first federal law to mention deepfakes, requiring a report on the implications of deepfakes at the national security level and the possibility of foreign countries using it for malicious activities. It also includes researching counter technologies that can be used by the government against foreign deepfakes. However, its focus is on foreign affairs, as the purpose is to delegate military activities, resulting in insufficient action against domestic issues with deepfakes. Another bill is the DEEPFAKES Accountability Act (2021), which specifies that any “advanced technological false personation record with the intent to distribute such record over the internet” should have a digital watermark and a disclosure describing the contents that were altered. This bill is not sufficient as watermarks and disclosure notes can be removed without requiring much effort, so it does not improve the situation. Lastly, the Deepfake Task Force Act (2021) would temporarily establish the National Deepfake Provenance Task Force, whose goal is to investigate digital content forgeries, including deepfakes, and provide recommendations on how to reduce them. A major problem is that the task force is only temporary and will be terminated once they submit their report. As the technology is rapidly advancing, the government should at least take periodic updates on how digital content forgeries are being made and give timely responses to those changes. Apart from the fact that none of the bills that regulated deepfakes were passed as a law, the influence of these bills on malicious deepfakes are fairly weak in that the actions they propose are insufficient in comparison of what is needed.

Looking at the bills mentioned, none of them target the root problem: legally penalizing the creation of deepfakes for malicious intent. This is because the Supreme Court’s interpretation of the First Amendment would likely strike down any laws that regulate deepfakes as unconstitutional. As a better approach, the government could first determine the fine line between deepfakes that are protected under the First Amendment and those that are not to create a uniform standard. The purpose of this standard is to prevent infringement of the freedoms of speech and expression while also allowing the government to give strict punishment to people who create malicious or illegal deepfakes. Furthermore, the Supreme Court cannot rule it unconstitutional as the government regulations only apply to deepfakes that are not protected.

The Developers

The creators of deepfake technology failed to take ethical responsibility when releasing their products, allowing for the malicious usage of deepfakes. Most developers deny liability for user actions and leave it up to the user to decide how to use the software. Even worse, the software is sometimes advertised to encourage illegal actions, further highlighting the lack of effort from the creators to prevent harmful deepfakes.

In creating and releasing deepfake software, the developers do not incorporate effective measures to regulate authorized use of the software, believing that it is not their responsibility to do so. The popular rise of deepfakes started in 2017 with the Reddit user "deepfakes," who posted various pornographic videos on the platform (Cole, 2017). Samantha Cole is a skilled technology journalist with ten years of experience and focuses on topics such as AI and sex work. Despite the publisher, Motherboard by Vice, being an online magazine, Cole includes an interview from "deepfakes," revealing their personal opinion. When asked about the consideration of ethical issues, "deepfakes" says, "'Every technology can be used with bad motivations, and it's impossible to stop that … The main difference is how easy [it is] to do that by everyone. I don't think it's a bad thing for more average people [to] engage in machine learning research'" (Cole, 2017). "deepfakes" adopts the technological deterministic perspective, believing that nothing can be done to control how technology is being used; it is deterministic. However, they encourage more people to get involved with deepfake software. This is also evident in the release of their code and training data on GitHub, claiming that it is not "worth the trouble to keep it secret from everyone" (Wu, 2017). Furthermore, there is no mention of a disclaimer or ethical regulation in the code. Releasing an unregulated product to the public allows anyone to explore the unlimited possibilities, for better or worse. Although the intentions of "deepfakes" may be to let others uncover the various beneficial applications of deepfakes, it also leaves the software out in the open for those with malicious intent to exploit.

The denial of developer responsibility also plays a major role in the uncontrollable consequences of DeepNude, a software that essentially "undresses" photos of people to create fake nude pictures. Again, Cole wrote a series of articles thoroughly covering DeepNude with real time updates on the situation and obtained personal statements from the creator "Alberto." The continuation of Cole's work in pornographic deepfakes adds to the trustworthiness of her articles. Alberto justified the release of DeepNude by arguing that "what you can do with DeepNude, you can do it very well with Photoshop (after a few hours of tutorial)" (Cole, 2019a). But, it offers a simpler process and removes the need for specialization in photoshopping; it eliminates some of the barriers for creating fake pictures, giving easier access to creating pornographic deepfakes. Furthermore, there are few measures that restrict how users use the app. The produced photos have a large watermark, but that can be replaced with a small stamp in the corner after paying $50 to upgrade to the paid version, which can easily be cropped out or removed with Photoshop (Cole, 2019a). The freedom given by the app is possibly a reason for why it quickly went viral. In a followup article the next day, Cole (2019b) reported that Alberto "didn't want to be the one responsible for this technology … [he does not] want to be the ones to sell it." So, he hastily took it down but did nothing else, allowing for the spread of unauthorized copies of the software, such as the start of a Discord server that shared a revised version of

DeepNude (Cole, 2019c). The new version had no watermarks and was free from various bugs that made the original software frequently crash. Although the server was later banned by Discord, the restriction-free software is still in the hands of many people. The release of DeepNude created a ripple effect that continued to disseminate due to the poor choices on Alberto’s part.

To make the situation worse, deepfake software is sometimes presented to hint at the illegal applications of deepfakes. Facemega is an app that allows users to replace faces in videos. The app had an ad campaign on Meta products, such as Facebook and Instagram, and one of the ads contained sexual content with a deepfake of the celebrity Emma Watson (Tenbarge, 2023). Tenbarge (2023), a tech and culture reporter for the highly-credible NBC News, also commented that “some of the ads showed what looked like the beginning of pornographic videos with the well-known sound of the porn platform Pornhub’s intro track playing.” The suggestive nature of these ads encourages people interested in this type of content to use the software for their own entertainment at the expense of others. Furthermore, Tenbarge (2023) notices that the “terms of service for the app … say it does not allow users to impersonate others via their services or upload sexually explicit content.” This creates a contradiction as the ads suggest the users to do something that is unauthorized. The producers of Facemega should be the most well-versed in the policies of their software but encourage usage that violates the terms of services, which shows how they value the economic benefits over the legal consequences and ethical responsibilities. Apps like these already give easy access for generating fake, unethical content, and this kind of advertising will only push more people to use deepfakes for malicious purposes.

On the other hand, there are some developers who acknowledge the ethics of deepfakes and are actively restricting unethical usage. An example is FaceSwap, which originated from the code released by “deepfakes.” Since the creation of the repository in GitHub, FaceSwap has an ethics manifesto to list the unauthorized uses of the software, including creating inappropriate content or using someone’s image without consent (deepfakes, 2024). It also states, “We will take a zero tolerance approach to anyone using this software for any unethical purposes and will actively discourage any such uses.” This claim is expressed throughout their product: having a code of conduct on GitHub, adding the ethics manifesto to their website, regulating inappropriate content on their forum, etc. (*FaceSwap,* n.d.). It is clear that the developers want to ensure FaceSwap is used for ethical purposes. However, this is not sufficient. The regulations in place are only in effect on specific platforms that are under the developers’ control, such as the FaceSwap Discord and forum. Malicious users can still use the software privately for their own purposes without restrictions. The software does not cover the deepfakes with a watermark nor anything visible that notifies the audience it is fake. Even if the developers do include it, anyone has access to the code and can remove it on their own. Despite the attempts of the FaceSwap developers to regulate unauthorized use of their software, more action is needed to limit the possibilities the user can do with FaceSwap.

Unlike the previous examples, Synthesia is an established company, but their efforts to be ethical and regulate use of their deepfake products were still compromised. The company offers over 85 deepfake avatars that have varying appearances, and they can be manipulated to speak in 120 languages and accents (Satariano & Mozur, 2023). Satariano and Mozur, working under the widely-respected New York Times, report internationally about online disinformation. Synthesia “has a four-person team dedicated to preventing its deepfake technology from being used to create illicit content” (Satariano & Mozur, 2023). To create deepfakes of a specific person, regardless if it is a celebrity or uploaded by the user, the process is transparent and consensual

(*Ethics*, n.d.). Through their efforts, the company is actively trying to implement measures to prevent illegal misuse, and they seem to be effective on paper. But, it is difficult to review content created with its software and determine if it is harmful or inappropriate at a large scale. Thus, some deepfake avatars are uncaught and can be exploited by those with malicious intent. For instance, their software was used to create two avatars, "purportedly anchors for a news outlet called Wolf News," that were a part of a pro-China campaign (Satariano & Mozur, 2023). The goal was to use the deepfakes to increase the American support of China and spread misinformation to do so. This shows how the company is not the sole actor behind regulating deepfakes; Synthesia is putting effort into preventing unauthorized deepfakes, but it requires collaborative action from all relevant groups to have a large impact against malicious deepfakes.

The Creators

Although certain types of deepfakes are allowed under the First Amendment and should not have any legal restrictions, they can still have malicious consequences due to their deceiving nature. Thus, the creators should be responsible for clearly marking their creations as deepfakes, even if they have no intentions to harm others. A deepfake without a label, or with an unclear one, can result in people believing in the false content, and, by sending it to others, public exposure of the deceptive deepfake can rapidly increase, amplifying the effects of misinformation.

As mentioned before, the intent of the deepfake creators determine if their creations are protected by the First Amendment, so, under this definition, creators of harmless deepfakes are not a part of the actor network of malicious deepfakes. A prime example is the art project Big Dada by artists Bill Posters and Daniel Howe. Big Dada consists of five deepfake videos, which were posted on Instagram, with the goal of "[interrogating] the power of computational propagandas," according to Posters (2019b). For each of the Instagram posts, Posters (2019a) added the hashtag #deepfake and a caption explaining that the deepfakes are part of Big Dada; the most popular of the five is a deepfake of Meta's CEO Mark Zuckerberg with about 125,000 views as of now. The video gained much attention from various news outlets and caused Meta to question "their internal policies regarding deep Fake videos and computational forms of propaganda that exist on their platforms" (Posters, 2019b). This indicates that Posters conveyed the message he wanted to get across through his creations. The deepfakes served only their intended purpose, and there was no harm done.

However, deepfakes are different from other forms of expression because they are created to depict a target doing something that never happened; they are created to mislead its viewers. Similar to Posters, Eliot Higgins (2023), the founder of the investigative collective Bellingcat, generated a satirical deepfake of Donald Trump and posted it on X with the caption, "Making pictures of Trump getting arrested while waiting for Trump's arrest." But, the outcomes of the two are very different: "some people [passed] them [deepfakes of Trump] off as genuine" (*AI-Generated Images*, 2023). The main difference between the two is in the captions. Posters directly labeled his posts as deepfakes, whereas Higgens only suggested, in the captions, that the images are created by him because he "assumed that people would realise Donald Trump has two legs, not three" (*AI-Generated Images*, 2023). Not having a clear caption introduced some ambiguity that possibly misled people into believing in the fake content. Even with the obvious illogical features, those who were deceived did not look closely and believed what they saw. Furthermore, some viewers shared only the photos on other platforms without any context, which likely intensified the confusion (*AI-Generated Images*, 2023). This example highlights the importance of labeling deepfakes as fake to prevent misinformation. Without it, even a harmless deepfake can have harmful effects on others, implying that creators do play a role in the malicious deepfakes network.

Conclusion

The growing presence of malicious deepfakes is largely due to the lack of government regulation, the ineffective control over unauthorized software use, and the absence of clear labeling by the creator. The government failed to enact effective legislation, and current laws allow for the continuation of deepfakes and their advancements. Moreover, most developers did not make ethics a priority when creating these technologies, making it easy for people with malicious intent to generate deepfakes. Lastly, regardless of their intentions, creators of deepfakes who do not clearly mark their creations as fake can fool others and create unintentional harm. From this analysis, we understand in detail how failures of different actors in taking action against malicious deepfakes led to the proliferation of them and what can be done to improve the situation. However, there are some actors that were not covered in depth, such as the Internet, which, as mentioned earlier, acts as an echo chamber to amplify the negative effects of deepfakes, and social media companies and their policies against fake and manipulated content. This poses further questions of to what extent is the advancement of deepfakes steered by the Internet and who is responsible for taking action against deepfakes: the social platforms where deepfakes are posted or the government.

References

AI-generated images of Trump being arrested circulate on social media. (2023, March 21). The Associated Press. https://apnews.com/article/fact-check-trump-NYPD-stormy-daniels-539393517762

Blitz, M. J. (2020). Deepfakes and other non-testimonial falsehoods: when is belief manipulation (not) first amendment speech?. *Yale Journal of Law and Technology*, *23*(1), 160-300. https://yjolt.org/sites/default/files/23_yale_j.l._tech._160_deepfakes_0.pdf

Cole, S. (2017, Dec 11). AI-Assisted Fake Porn Is Here and We're All Fucked. *Vice*. https://www.vice.com/en/article/gydydm/gal-gadot-fake-ai-porn

Cole, S. (2019a, Jun 26). This Horrifying App Undresses a Photo of Any Woman With a Single Click. *Vice* https://www.vice.com/en/article/kzm59x/deepnude-app-creates-fake-nudes-of-any-woman

Cole, S. (2019b, Jun 27). Creator of DeepNude, App That Undresses Photos of Women, Takes It Offline. *Vice* https://www.vice.com/en/article/qv7agw/deepnude-app-that-undresses-photos-of-women-takes-it-offline

Cole, S. (2019c, Jun 27). Discord Just Banned a Server Selling DeepNude, an App That Undresses Photos of Women. *Vice* https://www.vice.com/en/article/d3n9xa/discord-banned-a-server-selling-deepnude-an-app-that-undresses-photos-of-women

Contreras, B. (2024, February 8). Tougher AI policies could protect Taylor Swift--and everyone else--from deepfakes. *Scientific American.* https://www.scientificamerican.com/article/tougher-ai-policies-could-protect-taylor-swift-and-everyone-else-from-deepfakes/

DEEP FAKES Accountability Act, H.R.2395, 117th Cong. (2021). https://www.congress.gov/bill/117th-congress/house-bill/2395/text

Deepfake Task Force Act, S.2559, 117th Cong. (2021). https://www.congress.gov/bill/117th-congress/senate-bill/2559

Ethics. (n.d.). Synthesia. https://www.synthesia.io/ethics

FaceSwap. (n.d.). FaceSwap. https://faceswap.dev/

deepfakes. (2024). Faceswap [Software]. GitHub. https://github.com/deepfakes/faceswap

Gertz v. Robert Welch, Inc., 418 U.S. 323 (1974). https://supreme.justia.com/cases/federal/us/418/323/

Higgins, E. [@EliotHiggins]. (2023, Mar 20). *Making pictures of Trump getting arrested while waiting for Trump's arrest.* [Image attached] [Post]. X. https://twitter.com/EliotHiggins/status/1637927681734987777

Latour, B. (1996). On actor-network theory: A few clarifications. *Soziale Welt*, *47*(4), 369–381. http://www.jstor.org/stable/40878163

National Defense Authorization Act for Fiscal Year 2020, Pub. L. No. 116-92, 133 Stat. 1549 (2019). https://www.congress.gov/116/plaws/publ92/PLAW-116publ92.pdf

New York Times Co. v. Sullivan, 376 U.S. 254 (1964). https://supreme.justia.com/cases/federal/us/376/254/

Posters, B. [@bill_posters_uk]. (2019a, Jun. 7). *'Imagine this...' (2019) This deepfake moving image work is from the 'Big Dada' series, part of the 'Spectre' project. Where* [Video]. Instagram. https://www.instagram.com/p/ByaVigGFP2U/?utm_source=ig_embed&ig_rid=2b2f8203-eb13-4120-a5af-d7c7591cab41

Posters, B. (2019b, June 20). *The Zuckerberg Deepfake Heard Around The World.* Bill Posters. https://billposters.ch/the-zuckerberg-deepfake-heard-around-the-world/

Satariano, A., & Mozur, P. (2023, February 7). The People Onscreen Are Fake. The Disinformation Is Real. *The New York Times*. https://www.nytimes.com/2023/02/07/technology/artificial-intelligence-training-deepfake.html

Stone, G., & Volokh, E. (n.d.). *Freedom of Speech and the Press*. National Constitution Center. https://constitutioncenter.org/the-constitution/amendments/amendment-i/interpretations/266

Tenbarge, K. (2023, March 7). *Hundreds of sexual deepfake ads using Emma Watson's face ran on Facebook and Instagram in the last two days*. NBC News. https://www.nbcnews.com/tech/social-media/emma-watson-deep-fake-scarlett-johansson-face-swap-app-rcna73624

Weatherbed, J. (2024, January 25). *Trolls have flooded X with graphic Taylor Swift AI fakes.* The Verge. https://www.theverge.com/2024/1/25/24050334/x-twitter-taylor-swift-ai-fake-images-trending

Wu, J. (2017). deepfakes_faceswap [Software]. GitHub. https://github.com/joshua-wu/deepfakes_faceswap