Dhvani Patel, Senuri Rupasinghe

Professor Kulikowski

CS405

4 May 2024

The Deepfake Problem: The Problem, Consequences, and Potential Solutions

The idea for this project was inspired by the prevalence of image manipulation technology that superimposes the face of a real or imaginary person onto that of another, which is commonly known and will hereby be referred to as "deepfake." In recent years, deepfakes have become ever popular as a form of meme in many internet subcultures, due to the advancement and accessibility of generative artificial intelligence technology. There have been many companies developing artificial intelligence algorithms, which not only made the process of creating deepfakes much easier, but also lowered the barrier to obtain such technology. Because of these factors, we are now seeing an exponential increase in the creation and dissemination of deepfakes—from 2019 to 2023 one study found a 550% increase in the number of deepfake videos online, not including still photo deepfakes (homesecurityheroes.com). Due to the sheer volume of deepfake activity and unfettered public access via the Internet, there are bound to be cases where deepfakes are used to maliciously target a person and depict them in hurtful ways. The goal of our paper is to identify the personal, social, and political problems that result from the abuse of deepfakes and explore technological and legislative solutions to combat misuse.

Before delving into the problems with deepfakes, it's important to acknowledge that there is creative potential and benefit in the use of deepfakes, and the ability to make a deepfake does not always imply that the deepfake is being used for a harmful purpose. For example, there exist

apps that employ deepfake technology for the purpose of photo editing, and allow the user to alter their hair style, outfits, eyebrow shape, or see what they would look like as a younger or older version of themself, by combining photos from a pre-existing source. Deepfakes of this nature may not harm anyone as in such cases the person whose image was manipulated is usually the one creating the deepfake and thus consenting to the use of their image in the creation of the deepfake. At the same time, there are cases such as deepfake pornography of real people in which the use of deepfake is a clear unconsented violation of someone else's integrity, as well as many less severe examples that are in a gray area– and so it's not always obvious when a deepfake is or isn't a violation of someone else's liberty. As a result, it's necessary to develop ethical guidelines that are widely agreed upon and legal guidelines that are standardizable to determine when it is and is not appropriate to use deepfakes.

To figure out what makes a deepfake ethical, it's helpful to first introduce some terms to talk about deepfakes themselves. Deepfakes can be described in terms of a division between donor material and target material (dhs.gov). Donor refers to the material from which a face was taken and superimposed onto another photo, or audio was taken and modified to make another person seem to be saying something they never actually spoke. Target refers to the underlying video that is used as a base to alter upon. In certain cases, the face of the target is replaced with a donor, but their body is kept; in other cases, the target's face is kept but distorted to accommodate audio from the donor material.

It can be argued that there are two main components that determine whether the creation of a deepfake was ethical. Firstly, the donor material must be used with the knowledge and consent of any people within it, given beforehand. Second, the target material must be used with the same condition, that the people in it are aware of and consent to the use of their image. In the

case of celebrity pictures being used as donor material for sexually explicit photos, such as the deepfakes created of Taylor Swift earlier this year, creation would be unethical as the donor never consented to the use of their images in the deepfake. With such examples it is relatively obvious to understand that the subjects of sexually explicit deepfakes do not know about and do not want deepfakes of them being distributed without their consent.

In the case of self made deepfakes created in photo editors, if the developer chose to source all target images of hairstyles, outfits, and facial features from aware and consenting friends, then they would have upheld their share of the responsibility in creating ethical deepfakes. Yet it is also true that creating ethical deepfakes relies on users to source their target material responsibly. In an ideal case, the individual decides for themselves whether or not to use their face as donor material. However, there is a problem here that has caused many other challenges in regulating software: maintaining a standard of ethics relies on both parties to act responsibly, though neither has full control over the other.

In practice, this is not guaranteed. The government can require that developers adhere to certain protective regulations when creating apps, such as banning non-consensual deepfakes– but developers have little way of knowing whether images fed to their software by users are ethically sourced. Similar issues have occurred when regulation attempts to enforce age limits on apps, as developers can ask users whether they are over 18– but have no choice other than to trust self reported information which will likely contain bias. Thus, the goal of preventing certain behavior such as unethical deepfakes or underage users is never *truly* solved but the company or developer can absolve themselves of liability from the problem. Furthermore, due to the decentralized nature of the internet, there will always be avenues for piracy and illegal software that circumvents legislation. The situation can be analogized to Pandora's box—once someone

has created unfettered deepfake software somewhere and it spreads, it is virtually impossible to retroactively prevent access to it.

There are also scenarios when deepfakes have been made of people without their consent where the ruling is less obvious. Many deepfakes of political figures have been created without the consent of the subject as a form of caricature or satire, which has historically been protected as freedom of speech. In one such case, a deepfake is created of a scene from the hit show *Breaking Bad* where character Saul Goodman teaches another character Jesse Pinkman that he must hide his illegally obtained wealth by money laundering. However, the face of Saul Goodman is deepfaked as Donald Trump and Jesse Pinkman is deepfaked as Trump's son in law, Jared Kushner. This deepfake was likely created without the consent of Trump or Kushner, likely also disparaging their character, but seems less egregious than the cases of Taylor Swift.

This is because the creator of the video does not try to pass off the deepfake as reality. Trump and Kushner's faces are used and recognizable but the characters Saul and Jesse remain largely unmodified, still having the same hair and outfits, and the author has clearly labeled the video as a deepfake in the title. This is the same reason that political cartoons, caricatures, and performances like SNL are able to perform satire—in each medium, though there is resemblance to real people, viewers can easily distinguish the form of art from reality. Art is two dimensional and often cartoonish, and actors can imitate but never quite recreate a real person whereas deepfakes can.

Perhaps it is not definitive enough to conclude all deepfakes made without the consent of the subject are unethical. The difference is in intent: in this case, an intent to pass the deepfake off as reality or leave it as ambiguous was not present. However, in the Taylor Swift case, deepfakes were not explicitly labeled as fake and also of a predatory nature. Deepfakes created

of a widely known personality for the purpose of publicly calling out their actions or even mocking them (outside of bullying or hate speech) can be regarded as a first amendment matter, whereas deepfakes created with discriminatory intent or the purpose of sexual harassment are more closely scrutinized.

This observation also falls in line with existing guidelines regarding establishing defamation, suggesting the laws about defamation can be legal grounds upon which victims of deepfakes can receive retribution. Parodies and criticisms of public figures are protected speech under first amendment laws and so the legal concept of defamation does not apply to the Trump deepfake. However, acting with "actual malice" or "reckless disregard for the truth" with harm to another's reputation can qualify as grounds for defamation and so deepfakes of a derogatory nature like in the Taylor Swift example suit the definition.

Though these examples show how deepfakes have a significant personal impact, the misuse of deepfakes can also have wider reaching consequences for society, if used to spread misinformation. Deepfakes of political leaders created with an intent to pose as reality or spread false information can be used to artificially manipulate public opinion on a societal level. We have already seen that large social media platforms such as Facebook have the ability to influence public perception about matters such as elections. They selectively adjust the social media content a user sees to be biased one way or another due to confirmation bias in human behavior and the tendency to agree with opinions one is exposed to over and over again. Deepfakes have the potential to *amplify* this behavior in individuals participating in echo chambers, as they do not question content that aligns with their biases and are more vulnerable to falsified information. Laws should be passed that prevent deepfakes with an intent to be passed off as reality and spread misinformation from being published.

Furthermore, the ideas of copyright and licensing can also be used to combat the misuse of deepfakes. There is arguably a third principle worth mentioning: for a deepfake to be truly ethical, any artificial intelligence models such as the encoder and decoder neural networks used to create deepfakes should be trained on data that is consensually sourced. This is very hard to achieve given that big tech companies have created a landscape in which it is normalized to prevent consumers from having power or ownership over their data. Consumers often don't know how their data is handled, what it's used for, or who is in possession of it. Creating more legislation to protect consumer privacy and require companies to be transparent about the data they source to train their artificial intelligence models can lead to more ethical deepfake models. The concept of licensing is similar in that an individual can choose to distribute media while retaining some control over who gets to see or use their creation and what it can be used for. Just as people can apply different licenses to their art and images to determine whether others may use it for profit, a similar category of license should be created, coupled with AI transparency laws, to allow people to specify whether or not they consent to the use of their media as training material for models.

In the end, efforts to curb the misuse of deepfakes should focus less on how to stop the creation of unethical deepfakes, and place more accountability on users of such software to conduct themselves ethically. Legislation may be able to mitigate the misuse of deepfakes by providing retribution to those already affected, and establishing the legal precedence that there are punishments to deepfake misuse. Over time this will change public perception of deepfakes from the idea that AI technology is a wild west with no consequences, which will disincentivize the simplest of deepfake related threats.

It is also worth making tech companies that are responsible for the sharing of information, like Google or certain social media giants, comply to certain consumer protection laws which can include protections against misuse of deepfakes. Working together with companies can prevent unethical deepfakes from being spread by stopping them at the source with deepfake detection and other technological solutions.

Subsection: Technological and Curative Solutions

Forensic Analysis

To begin discussion on the technological and curative solutions of detecting deepfakes, let us introduce the concept of forensic analysis. Constructing a deepfaked face involves digital software, and that means the software of the deepfake will most likely leave digital artifacts in the image or video. "Digital artifact" in this context means the noticeable attributes of the image or video that seem out of place and inconsistent (DSCI, 2024). Examples of digital footprints would be inconsistencies in shadows, lighting, and reflections. Specific software to detect these inconsistencies are available and should be made more accessible to society. A specific way to catch these irregularities is to use specific deep learning models.

Deep Learning Solutions

Using deep learning, we can use Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to detect deepfakes. CNNs are most commonly used for video and image analysis tasks. We could train a CNN to classify an image or video as generated by a deepfake algorithm, or otherwise. There are unique elements and features to a deepfake that make it stand out, and we could train our network to catch these inadequacies. Another possible

solution could be the use of Recurrent Neural Networks (RNNs). RNNs are unique because they can be used for the *temporal* analysis of videos. They are able to comb through a video, frame by frame to catch inadequacies in the flow of the video, as deepfakes are known to have issues in mimicking the natural flow of human movement. For both CNNs and RNNs, the technical knowhow gets quite intense as there is a significant amount of deep learning involved, so let figures 1.1 and 1.2 be a sufficient overview of the processes involved. For the purposes of this paper, we will understand that they exist, and are very promising solutions to the infamous deepfake problem (DSCI, 2024).

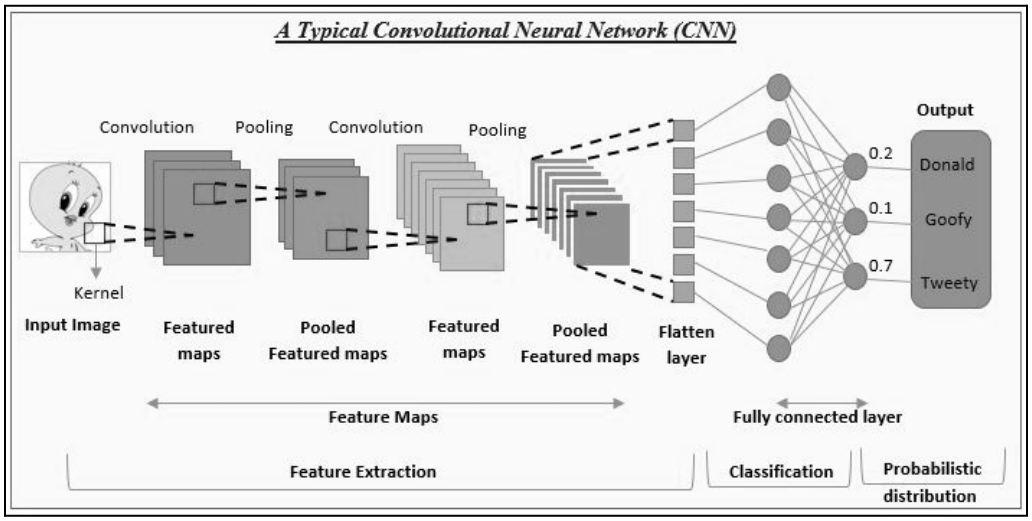

Figure 1.1 Convolutional Neural Network Structure (Shah, 2022)

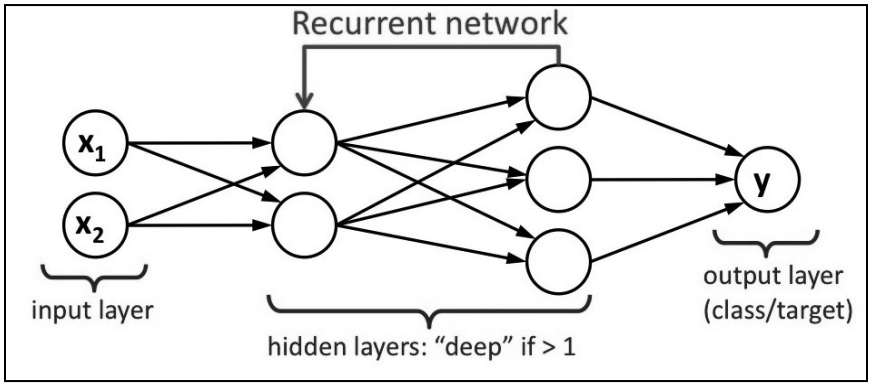

Figure 1.2 Recurrent Neural Network Structure (Dozmorov, 2020)

Watermarking Solutions

Another curative solution to deepfake detection involves watermarking. In the practical scenario where an ordinary person is viewing a video, there must be something to let that viewer know that what they are currently viewing is deepfaked. For them not to know would be a violation of trust. A simple solution would be to mandate a law such that every generated deepfake video is required to have a watermark on the bottom right corner of the screen, for example. This way viewers would have the complete knowledge that what they are currently watching is fake. For generative AI to not do this, should be unlawful. Companies that produce generative AI technology should take responsibility and provide the user with full disclosure of the information and images its AI produces.

Increase in Education and Awareness Solutions

Another rather obvious solution to the deepfake problem is increasing education and awareness on the subject of deepfakes and AI as a whole. Current research conducted by the Pew Research Center, suggests there is a distinct qualifier that suggests whether a person may demonstrate greater awareness of AI in daily life; that qualifier is those individuals that have higher levels of education and income (Figure 1.3). This is important to note as it implies that the *social* and *economic* standing of an individual is expected to determine whether they may understand the full complexity of AI. Therefore, there must be efforts in increasing awareness of AI for those that may not have higher levels of education and income.

At the rate it is currently excelling at, there is no doubt that AI will be a normal part of daily life in the future. This advent of monumental technology is no different than the major jumps in advancement we've had in technology in the past. For every introduction of major new technology, there is the resulting educational and awareness accompaniment. Internet and digital

literacy, print media and literacy, electricity and electrical safety, automobiles and road safety. The introduction of AI should be no different. Concrete ways to introduce awareness would be to hold community town halls or outreach events so that the general public, regardless of income status or education level, would be able to learn more about AI and the danger of deepfakes. Another simple way to raise awareness would be for the many companies that produce such AI to take responsibility and run either commercials or ads that spread the word about how deepfakes work. In this way, we can again see how the companies must take responsibility for the content generated by their AI.

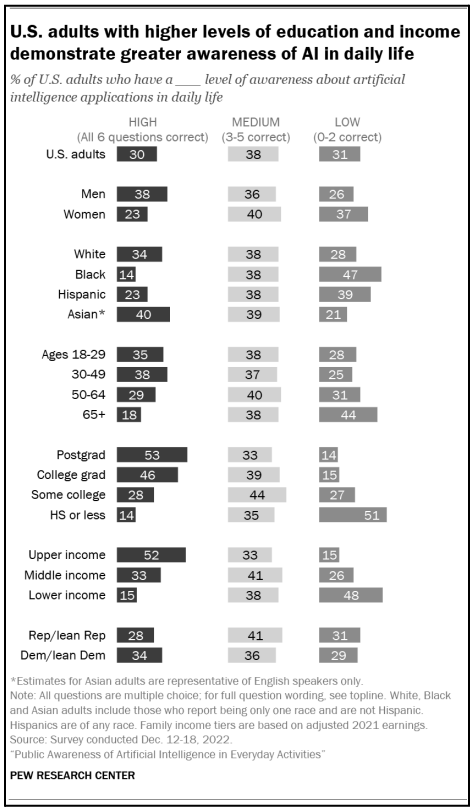

Figure 1.3 U.S. adults with higher levels of education and income demonstrate greater awareness of AI in daily life (Pew Research Center, 2023)

Conclusion: The Takeaway

At the end of the day, deepfakes are a major side effect of the advent of generative artificial intelligence. As AI continues to improve, the trickiness and the deceptiveness of deepfakes will increase, and detecting them will become more difficult. This is why it is imperative that solutions to the detection of deepfakes are implemented and put to use, so as to keep up with the continuing advancements of AI. We note that the societal implications of deepfaked technology is highly consequential. As mentioned in the educational awareness solutions, social and economic standing are indicators of awareness of AI, making AI yet another way to divide and separate our society further. We would further like to conclude by mentioning that the solutions proposed and highlighted here are quite practical and efficient. There are concrete steps that can be taken so as to implement these solutions. Those in charge of legislation, companies, and business must do their part in making these solutions a greater part of our reality. Artificial intelligence is simply a new technology that is opening worlds and opportunities to learn and grow as a society. However, we must acknowledge the dangerous side of AI and the possible consequences for society. Deepfaked images and videos are at the top of this list.

Our paper is centered around the fact that the recent development of deepfake technology presents a variety of new societal problems that fail to be addressed by our current justice system. Deepfakes can be used to maliciously target a person by depicting them in hurtful, nonconsensual ways, being used as collateral for extortion, or being spread to a bigger audience to humiliate or otherwise damage the social perception of that person. Additionally, the mere fact that deepfakes can exist calls for change in our perception of truth and the process of determining truth and reality from falsehood, which is always a step behind the pace of technological

development. Video evidence can now be faked and is not a surefire form of proof. This should be reflected in the way our courts and political landscape operate, as both will be affected by an increase in fraud and misrepresentation. Our online culture of being unwarily open with one's digital footprint now leaves the average person more vulnerable than ever to personalized deepfake attacks as anyone with a digital footprint of photos can be taken, edited, and used in a deepfaked video or image. This can happen to any ordinary person. All this takes is for their image to be on some digital platform, some type of social media or have a simple online presence. Given the consequences described above, it is imperative that protections for consumers are put in place as soon as possible.

The advent of generative artificial intelligence is, while posing many benefits, poses a great many threats to society as well. In this paper we have defined explicitly the extent to which a deepfake can affect any individual on a personal level. We have also provided legislative solutions to provide deepfake incidents with legal precedence, and technological and curative solutions to effectively detect deepfaked videos or images. By discussing such solutions and the rippling societal effects of deepfakes, we conclude that the deepfake problem must be attended to with haste and importance.

We will very soon reach a deepfake singularity. Deepfaking technology has developed to such a point that deepfakes are between often and always indistinguishable from reality. We as individuals can no longer rely on our innate ability to be certain of or pick apart the information given to us, or look for telltale signs of artificially generated information. Public perception of reality must change: "What is your perception of reality? Is it the ability to capture, process, and make sense of the information our senses receive? If you can see, hear, taste, or smell something, does that make it real? Or is it simply the ability to feel?" (A Deepfake Singularity 0:15-0:35)

Works Cited

Dozmorov, Mikhail. “Day 5: Generative Adversarial Networks, Autoencoders, Recurrent Neural Networks, LSTM, GRU, Sequence Learning.” Deep Learning with R, 12 June 2020, bios691-deep-learning-r.netlify.app/class/05-class/.

DSCI. “Unmasking The False: Advanced Tools And Techniques For Deepfake Detection.” Cybersecurity Centre of Excellence (CCoE), ccoe.dsci.in/blog/deepfake-detection#:~:text=Analysing%20Digital%20Footprints,during%20the%20deepfake%20creation%20process. Accessed 3 May 2024.

Foley, Joe. “23 of the Best Deepfake Examples That Terrified and Amused the Internet.” Creative Bloq, Creative Bloq, 31 Dec. 2022, www.creativebloq.com/features/deepfake-examples.

“Increasing Threat of Deepfake Identities.” *Homeland Security*, www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf.

Kennedy, Brian. “Public Awareness of Artificial Intelligence in Everyday Activities.” Pew Research Center, Pew Research Center, 15 Feb. 2023, www.pewresearch.org/science/2023/02/15/public-awareness-of-artificial-intelligence-in-everyday-activities/.

Public Broadcasting Service. (n.d.). *Defamation*. PBS. https://www.pbs.org/standards/media-law-101/defamation/

Shah, Saily. “Convolutional Neural Network: An Overview.” Analytics Vidhya, 15 Mar. 2022, www.analyticsvidhya.com/blog/2022/01/convolutional-neural-network-an-overview/.

"State-of-Deepfake-Infographic-2023.Pdf." *Home Security Heroes*, 2023, www.homesecurityheroes.com/state-of-deepfakes/assets/pdf/state-of-deepfake-infographic-2023.pdf.

"This Is Not Morgan Freeman - A Deepfake Singularity." *YouTube*, YouTube, 7 July 2021, www.youtube.com/watch?v=oxXpB9pSETo.

YouTube. (2019, September 18). *Better Call Trump: Money laundering 101 [deepfake]*. YouTube. https://www.youtube.com/watch?v=Ho9h0ouemWQ