The Technical Implications and Societal Impacts of 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

Ian Switzer

Spring 2023

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

Richard D. Jacques
Department of Engineering and Society

Introduction

The twenty-first century has brought countless technological developments, but one of the most influential has been the advancement of artificial intelligence (AI), which has made its way into everything from cars to mobile phones. Machine learning has become a viable strategy for generating media such as text, audio, images, and videos, with a massive increase in publicity in the past year. AI-generated text, though impressive, hardly raises any large-scale ethical concerns by itself. The main ethical concerns arise from the training data used for these models. On the other hand, AI-generated audio clips, images, or videos, also known as deepfakes, can potentially have huge impacts on society as a whole.

Research in the late 90s and early 2000s paved the way for the advanced deepfakes that exist today. Sophisticated AI can produce fake audio clips of someone talking that sound indistinguishable from a real person to the naked ear. That audio can be fed into another algorithm that generates a talking face that matches the words. This video could be a deepfake of anyone, including celebrities and politicians. Even the most advanced detection methods struggle to catch everything, and creation is advancing much faster than detection. Though deepfakes have some potentially beneficial uses, most deepfake technology has been used for nefarious purposes, such as to create explicit videos of celebrities or blackmail politicians. Advanced deepfakes have only been around for a few years, but the technology has already progressed to the point where anyone can create them with relative ease. This technical discussion will seek to analyze the different methods for creating deepfakes and determine how deepfakes can be recognized at a technical level. This STS discussion will attempt to determine the potential societal impacts of deepfakes and how online platforms can respond to sophisticated misinformation effectively.

Technical Discussion

Fake images and videos can be created manually using editing tools like Photoshop, but the term ‘deepfake’ is a combination of deep learning and fake, and it refers to images and videos specifically created by machine learning software. Deepfakes are created using deep learning, which is an advanced form of artificial intelligence. One common method for making deepfakes is using Generative Adversarial Networks, which involves two neural networks that work against one another. The first neural network attempts to generate an image or video based on a sample dataset, and the second neural network guesses whether the generated image or video is fake (Goodfellow et al., 2014). This process repeats until the second neural network can no longer determine whether the generated media is fake. At that point, the program decides that the output is good enough and presents it to the user. The sample dataset can contain photos or videos of celebrities, politicians, or anyone else, and the result will be a deepfake of that person. Another method for creating deepfakes uses autoencoders, which use encoding and decoding algorithms to splice the face from one person onto another (Nguyen et al., 2019). This can be used to generate a video of one person speaking based on a video of someone else saying the same thing.

Researchers at the University of Washington used neural networks to create a deepfake video of Obama based on an audio clip (Suwajanakorn et al., 2017). Their technique involved generating mouth shapes that fit the audio clip and then texturing it to fit the target face. These are just a few examples of methods for creating deepfakes. Machine learning is always evolving and offers a myriad of techniques for generating entire fake images and videos, splicing one face onto another, creating lip and mouth shapes from audio clips, etc. Advances in machine learning have allowed these algorithms to be run on any computer, regardless of hardware, but with the

rise in cloud computing technology, huge machine learning models have been moved to the cloud to allow anyone to access them from anywhere.

The cloud has become a computing hub for all kinds of artificial intelligence. Industry-leading AI companies such as OpenAI are hosting websites where users can use their AI software free-of-charge and without needing to run it on their own hardware. The combination of accelerated hardware capabilities and increased efficiency via cloud computing has allowed AI programs like ChatGPT to spread world-wide in a matter of weeks. ChatGPT, created by OpenAI and released to the public in December 2022, reached 100 million monthly users in just two months, a feat which took TikTok nine months and Instagram two and a half years (Cerullo, 2023). It is by far the fastest-growing app in history. Though the initial release of ChatGPT only handled text, the newest version, released in March 2023, can accept documents, images, and other forms of input side-by-side with text and analyze them according to the user's directions. It may only be a matter of time until ChatGPT can respond with audio, images, and videos. ChatGPT is also not the only machine learning algorithm that is open to the public. There are dozens of sites on the internet that offer both free and paid ways to create fake media (Gaur, 2023). If one were to use a combination of audio generation, face swapping, and image-to-video services, they could easily turn a picture of someone into a video of them saying anything.

As deepfakes have grown increasingly common, researchers have also focused on identification techniques to determine whether images and videos are real or fake. Early deepfakes that used face-swapping were easily detectable because the face either never blinked or did not blink at a realistic frequency (Li et al., 2018). However, soon after this research was published, deepfake synthesis algorithms were improved to include blinking. Similar to a lack of blinking, deepfake faces often have slightly different-colored irises, and deepfake models often

estimate light levels incorrectly across the face. Deepfake generation models also struggle to portray teeth correctly. These inconsistencies are known as face-warping artifacts, and are detectable by neural networks (Gaur, 2023). Another technique, developed in 2019, uses machine learning to evaluate a person's consistent facial expressions based on known authentic videos and compares it to the facial structures in a questionable video to see if these facial structures change unexpectedly (Agarwal et al., 2019). This approach has several limitations, the most significant being the reliance on existing videos for training. This model would be impractical for anyone besides famous figures for whom there is ample video footage. The model also falls short when the subject is in varied contexts, such as looking at the camera versus looking at someone else while talking, though this limitation is largely based on the variation present in the dataset. A third method for detecting deepfakes involves training a neural network to analyze inconsistencies between frames in videos because the GAN method for creating deepfake videos is unaware of the last frame it created when generating the next one (Guera & Delp, 2018). This technique is much more robust because it relies only on the video in question itself for detection, rather than a large pre-existing dataset for a particular person.

Several large companies have recognized the growing importance of deepfake detection and have contributed data towards research. For instance, Google created the DeepFake Detection (DFD) dataset, which contains 100 real videos produced using paid actors and over 3000 deepfake videos based on the real ones (Gaur, 2023). Facebook produced the DeepFake Detection Challenge (DFDC) dataset, with more than 5000 videos. Datasets like these can be used to train detection algorithms on deepfakes that are comparable to those found on the internet. It is critical that the detection algorithms stay up-to-date on the latest nuances because deepfake technology is rapidly evolving past anything that has been seen before. It is nearly

impossible to predict the advancements that the next generation of deepfake technology will bring. Governments and other influential entities are aware of the existence of deepfakes and recognize the importance for detection and prevention, but they will need to be hyper-aware going forward because the world is on the verge of an AI-centered industrial revolution (Bowles, 2020). The current state of technology will be overturned in a matter of decades, according to Bowles, with AI effectively wiping out middle-class jobs. With such a bleak outlook, everyone needs to be cognizant of the role AI plays in their daily lives, which includes social media and false information.

STS Discussion

There are countless applications for deepfakes that are indistinguishable from real images and videos, some good and many worse. Machine learning has been used for special effects in movies to simplify the generation of computer graphics (Miller, 2022). Deepfake audio has been used to re-dub movie lines without the need to re-record. Deepfake technology also has the potential to be used to assist hearing-impaired people by presenting a real-time synthetic video for lip-reading based on phone call audio (Suwajanakorn et al., 2017). Deepfakes also have the ability to enhance education. Lectures could be much more compelling and memorable if augmented with deepfake videos of historical figures, as Loveleen discusses in her book on deepfakes (Loveleen, 2023). The book also notes the prospect of deepfakes being used by activists and reporters to maintain anonymity on the internet. Human rights advocates may feel more inclined to speak out on major social media platforms if they are able to do so anonymously, which is due in part to the difficulty in tracing a deepfake. There are many upsides to deepfakes, as long as the users of the technology are held to high ethical standards.

However, the downsides of deepfakes are much more important to consider. Deepfakes

can be used to imitate important political figures in order to spread misinformation or hate speech. On the other hand, prominent figures could attempt to discredit an incriminating video by calling it a deepfake. In Malaysia, someone leaked an explicit video of a senior Cabinet minister, which the Prime Minister said he believed was fake but was later authenticated by the Malaysian Cybersecurity team (McCoy, 2019).

Deepfakes also have the potential to ruin an individual's reputation and personal life by creating fake videos or audio that make them appear to be engaging in illegal or immoral activities. For instance, deepfakes can be used to create fake pornographic videos of individuals, which can be shared online without their consent. This can lead to significant personal trauma and emotional distress, as well as potential social and economic consequences. Moreover, deepfakes can also be used for blackmail, as individuals can be threatened with the release of fake videos or audio that can harm their personal and professional reputation. It can be difficult for individuals to fight back against blackmail via deepfakes because there are not a lot of laws regarding deepfake technology, and once something is on the internet, it is nearly impossible to remove it. The creation and dissemination of deepfakes also raises several legal concerns. For instance, deepfakes can violate an individual's right to privacy, identity theft, and defamation. The use of deepfakes can also result in copyright infringement, as the use of copyrighted material may be used without permission. All of these concerns stem from a technology that has seen very little regulation but has incredible potential for both harm and good.

In addition to countless ethical concerns over usage, deepfake technology also brings up questions about ownership. GANs and Stable Diffusion models have been shown to occasionally output images that are nearly exact copies of images from the training dataset, as well as images that are simple combinations of backgrounds and foregrounds from images in the training set

(Chen, 2023). This brings up the question of whether an image produced by stable diffusion can be copyrighted if there is a chance it came straight from the training set. Google's Imagen, a text-to-image diffusion model, has also been shown to leak photos of real people and copyrighted images. AI companies must be extremely careful with how they train their models, otherwise they will face countless lawsuits in the coming years.

Although the creation of deepfakes, both explicit and not, is nothing new, the use in political blackmail and propaganda just began to take hold a few years ago. Back in 2016, during the presidential election, it would have been unheard of for a candidate to claim that a recording of them was a fake. Nowadays, as Galston puts it, "a well-timed forgery could tip an election" (Galston, 2020). When a big story drops, such as an explicit image of a candidate getting leaked, that initial story usually gets the most coverage. Any updates, such as research revealing that the image is a deepfake, is often overlooked, or is less memorable than the initial story. This unfortunate pattern lets people with bad intentions use deepfakes to harass people, even if the deepfake is easily detectable, because people care more about the story itself than verifying it. In order to combat deepfake technology, the Defense Advanced Research Projects Agency (DARPA) office of the U.S. Department of Defense created a Semantic Forensics (SemaFor) program to analyze the semantic features of deepfakes (Defense, 2019). This program focuses on semantic errors such as mismatched earrings that existing detection algorithms would not notice. Semantic errors may be noticeable to humans, but it could be difficult to train a machine learning model to recognize them. A reliance on human efforts to identify deepfakes is both good and bad. Human review may produce less false positives and false negatives, but at the cost of a much slower pace than an algorithm where the speed is entirely dependent upon the hardware. Programs like SemaFor are useful but need to keep these drawbacks in mind and may want to

focus their human attention on only the most important cases.

Governments and social media platforms need to be aware of potential deepfakes because fake news can spread faster than the truth. It can be difficult for people to recognize deepfakes when they are not aware that they exist. To raise awareness and encourage innovation, Facebook announced a deepfake detection competition in 2019 with $10 million in prizes and grants and released a dataset of faces from consenting adults (Cole, 2019). Facebook has faced a lot of scrutiny in the past for their lack of content management policies and their unauthorized usage of user data. This competition is a step in the right direction for the company. Platforms like Facebook allow misinformation to run rampant, especially because most users are not technologically savvy. The ones that are tech savvy have the potential to wreak havoc on communities and governments. One could argue that misinformation is difficult to detect with an algorithm, but for a company as large as Facebook, they should have a rigorous combination of both human review and machine learning analysis. New threats are emerging constantly, and social media is at the forefront of the battle against misinformation.

Most innovative technologies begin in a phase where only experts can use them and then slowly progress towards a version that anyone can use. Some examples include computers, which were initially massive machines used by companies and researchers; mobile phones, which started out as big and bulky and difficult to transport; and the internet itself, which began as ARPANET, an academic research network funded by what is now DARPA (Lee, 2014). Machine learning recently transitioned into the phase where anyone can sit down at a computer and start building a deep learning model. Now, anyone can create simple deepfakes with relative ease. Sophisticated deepfakes require more advanced models but there are countless websites and apps that offer powerful deepfake functionality for cheap or free (Khan, 2022). Someone

could go onto their cell phone or computer, upload a photo or video, and have a brand new deepfake in minutes. People tend to trust pictures and videos that look completely realistic to the naked eye, which means that social media platforms need to be hyper-vigilant going forward because deepfakes will only become more prevalent and more advanced.

While a lot of responsibility rests with the social media companies that struggle to detect and remove deepfakes, the consumer must also be aware of what they are looking at. Some companies, such as DuckDuckGoose, have developed tools for the user to help combat misinformation in the form of deepfakes (DuckDuckGoose, 2023). They released a free browser extension that automatically scans any webpage the user visits and alerts them if it detects any deepfakes or manipulated images. Giving users the power to detect deepfakes is crucial if the general public is going to take a firm stance against ML misinformation in the future. Society cannot afford to rely solely on large corporations to change their practices for the greater good. It will take a combined effort from the governing bodies, the companies, and the public people. Educating the public about deepfakes is the first step and allowing them to perform their own analysis and detection is next.

Conclusion

The growing threat of deepfake technology has spurred some research into detection and prevention, but there is still a lot of work to be done to mitigate future problems. Methods for detecting fake pictures and videos have progressed quickly but have been constantly outpaced by developments in deepfaking techniques. Generative machine learning models have recently taken the public spotlight as models like ChatGPT and Midjourney force people to question the authenticity of digital artwork and written media. Artificial intelligence has reached the point where it can create art in the style of specific artists, generate fake, photorealistic images of

nature, and even write essays on books. Very soon, it will become extremely important to be able to detect AI-generated media with some level of accuracy. OpenAI's own AI-written text detector achieves only 26% accuracy (Aaronson et al., 2023). Social media giants are aware of deepfakes and have taken steps to prevent relevant misinformation but the fight against deepfakes is an uphill battle. Media forensics teams are forced to be reactive rather than proactive, reacting to new deepfake techniques as they appear. It is impossible to prevent deepfakes from being created or new techniques from being developed. The best course of action for governments and corporations alike is to remain vigilant and encourage people to question what they see on the internet. Companies like DuckDuckGoose are producing deepfake detection tools that are available to the general public, which will help people learn to do their own analysis when looking at things on the internet. In order to keep up with the lightning-fast development of AI, the public must be educated and aware of what can be seen on the internet. Social media giants are in a unique position where they can influence what people see on a global scale, so they are the ones that people will blame if false images or videos make their way across the internet. Unfortunately, it is becoming increasingly easy for the average person to create deepfakes, whether by using an app or website or constructing their own machine learning model. In the coming years, deepfakes will slowly break down public trust of pictures and videos on the internet, whether for good or bad.

References

Aaronson, S., Ahmad, L., Kirchner, J., Leike, J. (2023, January 31). *New AI classifier for indicating AI-written text*. OpenAI. Retrieved April 1, 2023, from https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text.

Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., & Li, H. (2019, June). Protecting world leaders against deep fakes. *IEEE International Conference on Computer Vision and Pattern Recognition 2019 Workshop on Media Forensics*. Retrieved October 28, 2022, from http://www.hao-li.com/publications/papers/cvpr2019workshopsPWLADF.pdf.

Bowles, C. (2020). *Future Ethics*. NowNext Press.

Cerullo, M. (2023, February 1). *ChatGPT is growing faster than TikTok*. CBSNews. Retrieved April 1, 2023, from https://www.cbsnews.com/news/chatgpt-chatbot-tiktok-ai-artificial-intelligence/.

Chen, C., Fu, J., Lyu, L. (2023, May 17). *A Pathway Towards Responsible AI Generated Content*. Retrieved April 1, 2023, from https://arxiv.org/pdf/2303.01325.pdf.

Cole, S. (2019, September 5). *Facebook just announced $10 million 'Deepfakes Detection Challenge'*. VICE. Retrieved October 28, 2022, from https://www.vice.com/en/article/8xwqp3/facebook-deepfake-detection-challenge-dataset.

Defense Advanced Research Projects Agency, & Corvey, W., DARPA (2019). Retrieved October 28, 2022, from https://www.darpa.mil/program/semantic-forensics.

DuckDuckGoose. (2023). Retrieved April 1, 2023, from https://www.duckduckgoose.ai/.

Galston, W. A. (2022, March 9). *Is seeing still believing? the deepfake challenge to truth in Politics*. Brookings. Retrieved October 29, 2022, from https://www.brookings.edu/research/is-seeing-still-believing-the-deepfake-challenge-to-tr

uth-in-politics/.

Gaur, L. (2023). *Deepfakes: Creation, Detection, and Impact*. CRC Press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. *Arxiv*. https://doi.org/1406.2661.

Guera, D., & Delp, E. J. (2018). Deepfake video detection using recurrent neural networks. *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. https://doi.org/10.1109/avss.2018.8639163.

Khan, I. H. (2022, February 13). *Top deepfake apps and websites you can try*. LinkedIn. Retrieved October 30, 2022, from https://www.linkedin.com/pulse/top-deepfake-apps-websites-you-can-try-imran-hussain-khan/.

Lee, T. B. (2014, June 16). *The internet, explained*. Vox. Retrieved October 29, 2022, from https://www.vox.com/2014/6/16/18076282/the-internet.

Li, Y., Chang, M.-C., & Lyu, S. (2018). In ICTU oculi: Exposing AI created fake videos by detecting eye blinking. *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*. https://doi.org/10.1109/wifs.2018.8630787.

McCoy, P. (2019, June 13). *I believe sex video is fake, says Mahathir*. The Straits Times. Retrieved October 27, 2022, from https://www.straitstimes.com/asia/se-asia/i-believe-sex-video-is-fake-says-mahathir.

Miller, T. (2022, March 21). *How deepfake technology is changing the movie industry*. Seat42F. Retrieved September 30, 2022, from https://seat42f.com/how-deepfake-technology-is-changing-the-movie-industry.

Nguyen, T. T., Nguyen, C. M., Nguyen, D. T., Nguyen, D. T., & Nahavandi, S. (2019). Deep

Learning for Deepfakes Creation and Detection. https://doi.org/1909.11573.

Suwajanakorn, S., Seitz, S. M., Kemelmacher-Shlizerman, I. (2017). Synthesizing Obama: Learning Lip Sync from Audio. https://dx.doi.org/10.1145/3072959.3073640.