Marlee Bryant

Dr. Monica Anderson

CS 565

27 November 2020

The Universal Threat of Deepfake Technology

The legitimacy of the phrase “Seeing is Believing” is being challenged by the emerging simplicity of modifying video and sound using modern technologies. Modern society relies on technology for access to news and information, with social media becoming a prominent platform for this information access. The intersection of fake content creation and the potential widespread effect of viral posts by individual, unverified creators presents a perfect pathway for the spread of fake news. Fake news is a concern on the minds of individuals and social media creators alike, with both parties focused on the detection and elimination of these nefarious attempts, but the integration of AI dramatically increases the complexity. Specifically, a new AI video editing technology allows for the creation of videos referred to as “deepfakes,” where the face of an individual replaces that of someone else in a video, making them appear to say and do things which they did not, producing shockingly realistic results. This editing software utilizes an unsupervised learning technique called a Generative Adversarial Network, where two neural networks work in unison, the Generator and the Discriminator (Yadav). The Generator analyzes user-provided images and integrates these into the output video, which the discriminator then evaluates for authenticity issues, performing this process reiteratively until the discriminator cannot detect that the video is inauthentic (Dixon, Yadav). While this software was intended for harmless purposes such as art and education, its potential for malicious manipulation abounds. Deepfake technology is one of the greatest threats in AI because of its far-reaching impacts in the

media and on the individual, especially due to its ease of accessibility, difficulty of regulation, and constant evolution.

Digital media, including social media, is one of the main sources for information regarding news, politics, and pop culture. While many are aware of the possibility of encountering fake stories and images, realistic deepfake videos are harder to identify and there have already been cases of these videos being believed and highly shared on social media platforms. In 2018 former president Obama spoke about the danger of deepfakes in a video that was itself a deepfake created by comedian Jordan Peele in an attempt to raise awareness (Greengard). With the subject of this video being an influential politician, it presents one of the key dangers of deepfakes, which can "create disputes in countries by influencing their election process by defaming the character of the politician" (Yadav). Defamation is not the only means by which deepfakes can influence political perception. Social media platforms utilize recommendation systems to gather the preference information of users and filter their feed to content which aligns with their preferences. This data can be used with deepfakes in a more targeted manner to influence opinion, for instance creating a video in favor of a politician "featuring a specific actor or leader that the user finds trustworthy" (Karnouskos). Deepfakes could be used in this manner even by seemingly reputable sources in an attempt towards profit-maximization (Karnouskos). While deepfakes focus on the creation of visual content, they can be coupled with software that synthesizes audio content to enhance believability. Existing AI-driven software can synthesize a person's voice with as little as one minute of original audio recording (Dixon). The effects of this software extend to businesses, where employees and shareholders can be manipulated, like when "crooks recently impersonated the voice of a U.K.-based energy company's CEO in order to convince workers to wire $243,000 cash to their account"

(Greengard). At face value, this is a concerning issue for anyone in the public eye, but even more daunting are the societal implications in regard to trust in the media. Deepfakes, especially in the context of fake news generation, "undermine the public confidence on what is seen, heard, and eventually believed to be true" (Karnouskos). In a society where digital media cannot be trusted, misinformation and skepticism will abound, leading to polarization and radicalization of opinions.

The origin of deepfakes was not in a news or political context, but instead pornographic. In 2017 a Reddit contributor used publicly available AI software to "impose the faces of celebrities onto the bodies of people in pornographic videos" (Dixon). Celebrity deepfake adult content is popular, abundant, and disproportionately affects female celebrities, leaving them feeling exposed and attacking their identity and moral standings through the depiction of explicit acts that they did not actually commit (Karnouskos). Celebrities are not the only target of this heinous act. One in 25 Americans has been the victim of the disgusting act of "revenge porn," where private or nonconsensual explicit content is released to the public by another individual (Karnouskos). This content can lead to public humiliation and difficulty finding employment, which is especially troubling when the video is a deepfake and the subject never actually performed the depicted acts. There already exists a plethora of deepfake pornographic content largely due to the ease of accessibility and use of deepfake software. The AI complexity of deepfake software "is hidden behind common easy-to-use tools and services that are available to the general public," making it simple for an average user to create incredibly realistic fake content (Karnouskos). In essence, this advanced technology with endless possibilities for weaponization has been placed in the hands of anyone with access to a computer.

The possibility for impact to the average person extends beyond the realm of adult

content. In legal proceedings videos are often used as evidence, so given their ability to be falsified they could "depict fake murders or frame a person for a crime he or she did not commit" (Greengard). Deepfakes could also be used to project a false identity in the hiring process, since available software can falsify CVs and "lets users create in real-time deepfake avatars for Skype and Zoom teleconference tools, simply from a photograph" (Karnouskos). Additionally, the reliance on AI-empowered cyberphysical systems could be manipuilated by deepfakes targeting the machines, for example a deepfake voice unlocking a self-driving car or projecting images to affect its sensors (Karnouskos). With so many possibilities for harm, one might wonder how deepfakes are being regulated and controlled. The Malicious Deep Fake Prohibition Act of 2018 imposes penalties on deepfake creators who intend to distribute these videos that "facilitate criminal or tortious conduct" (Greengard). In addition, the Accountability Act of 2019 describes actions which should be taken to prevent the spread of these videos (Karnouskos). Unfortunately, these laws lack perspective on the possibilities for deepfake videos and the mentality of creators, making them significantly less effective. For example, the Accountability Act requires "watermarks and clear labeling on deepfake content," a regulation which criminals would be unlikely to comply to. Another approach to controlling these videos is developing software to detect them, oftentimes the same kind of software which is used as a discriminator within deepfake systems. This exhibits an obvious concern, that "the technology to detect deepfakes may always be in catch-up mode" (Greengard). As deepfake technology continuously innovates, it will become more challenging to detect what is fake, and nefarious minds will imagine new and detrimental usages.

A publicly available powerful technology will always have the potential for misuse, especially when its negative effects were not considered and accounted for prior to its release.

Currently, legislation and counter efforts are struggling to catch up, and will only make effective progress if deepfake technology is thoroughly researched in order to make proactive efforts rather than reactive ones. On the individual level, awareness is paramount, and the development of good information analysis skills, for example checking the validity of a source and surveying a variety of sources, could foster less susceptibility to falsified news content (Greengard). While there is clearly a plethora of opportunities for misuse of deepfake technology, "the goal is not to stifle innovation," since this technology has potential for many positive uses as well (Greengard). Without proper research and awareness, the malicious usage of deepfake technology will continue to grow and evolve and could have detrimental impacts on society and on the individual.

Works Cited

Dixon, Herbert. “Deepfakes: More Frightening Than Photoshop on Steroids.” *Judges Journal*, Vol. 58, Issue 3, American Bar Association, 2019. https://web-b-ebscohost-com.libdata.lib.ua.edu/ehost/detail/detail?vid=11&sid=ad258bbe-569f-4764-af24-89ea94b0635e%40sessionmgr103&bdata=JnNpdGU9ZWhvc3QtbGl2ZSZzY29wZT1zaXRl#AN=138864171&db=aph

Greengard, Samuel. “Will Deepfakes Do Deep Damage?” *Communications of the ACM*, vol. 63, no. 1, pp. 17–19, ACM, 2020. https://web-a-ebscohost-com.libdata.lib.ua.edu/ehost/detail/detail?vid=3&sid=aa3ff6ae-7ca5-4176-9120-bc85164340f0%40sessionmgr4007&bdata=JnNpdGU9ZWhvc3QtbGl2ZSZzY29wZT1zaXRl#AN=141677481&db=aph

Karnouskos, Stamatis. "Artificial Intelligence in Digital Media: The Era of Deepfakes." *IEEE Transactions on Technology and Society*, vol. 1, no. 3, pp. 138-147, IEEE, Sept. 2020. https://ieeexplore-ieee-org.libdata.lib.ua.edu/document/9123958

Yadav, Digvijay and Salmani, Sakina. "Deepfake: A Survey on Facial Forgery Technique Using Generative Adversarial Network." *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, IEEE, 2019, pp. 852-857. https://ieeexplore-ieee-org.libdata.lib.ua.edu/document/9065881