"When Seeing Isn't Believing: College Students and Deepfake Detection."

Evan Mao

St. Edward's University

HONS 4349: Honors Senior Thesis

Dr. Billy Earnest

December 11, 2022

Abstract

Modern deepfakes (AI-generated video fabrications) first appeared in the 1990s. Since then, the technology has improved at such a rapid rate that, today, the ability to accurately distinguish real videos from artificially-generated ones is increasingly diminished. As a result, people from all walks of life are being fooled by deepfakes on a daily basis, often without ever knowing that they've been duped. To assess the depth of this societal-level deception, I showed four videos to college students and asked them to differentiate between deepfakes and authentic videos. Out of c. 40 participants the rate of successfully distinguishing real from fake was disappointingly low, suggesting that the technology has progressed so much that college students–presumably among the most technologically savvy users around–can no longer tell the difference. These findings serve as a warning sign that, in what is already an increasingly post-truth era, what has traditionally been a verifiable and reliable source of truth (i.e., video), is being lost as a basis of reality.

Introduction

Deepfakes are photos or videos that are created by a form of artificial intelligence in which a person in the photo or video is made to accurately look like someone else (Earnest, McGlone, Knapp, & Griffin, 2020, p. 493). This is a potential issue for society as a whole because deepfakes can literally alter our current perception of reality. When it comes down to proof or evidence in legal or complex situations, video evidence is usually the deciding factor of any outcome because it cannot be fabricated. However with the evolution and development of deepfakes, this notion of videos consistently being true is quickly transitioning to be false. In my research I will be looking to see the solutions to deepfakes and if people are capable of distinguishing the difference between deepfakes and unfabricated media.

Deepfakes have only been around for about five years, users made deepfakes of actresses like Scarlett Johansson, Daisy Ridley, Gal Gadot, Taylor Swift, and Emma Watson all for the sole purpose of making it appear that each actor was engaged in pornography (Earnest, McGlone, Knapp, & Griffin, 2020, p. 493). Gaining national attention at first, it was hard to identify that these videos were in fact false, and this is where the birth of deepfakes began. Users soon transitioned from producing fake pornography to creating deepfakes of other celebrities for comedic intent. They would alter the words of a celebrity by substituting it with someone else's voice (Earnest, McGlone, Knapp, & Griffin, 2020, p. 493)

There are many different forms of deepfake technology, but one way to create deepfakes is with the X2Face Method or First Order Motion technology. Đorđević, Milivojević, & Gavrovska (2020) explain how and why they created their deepfakes for their study by using the

technologies saying, “Both are based on the pix2pix network with changes introduced to its input and output layers. Driving and embedding networks differ by the presence of skip connection in the embedding network, and by the positioning and size of certain smaller inner layers, while the X2Face additionally uses two encoder-decoder networks, an embedding network and a driving network.” (p. 24.) There are some deepfakes that can be easily identified as fake, but others that really make a viewer question if it may be real or not. One deepfake phenomenon that swept social media in 2021, was an account on TikTok who viewers presumed to be Tom Cruise, but soon found out that each TikTok video was in fact a deepfake.

Deepfakes truly can alter our perception of reality by using political candidates or celebrities to spread misinformation. With the recent development of the Russia and Ukraine war, a deepfake video of Ukrainian President, Volodymyr Zelenskyy, was released on Twitter where it seemed he was surrendering to Russia, by telling his soldiers to “lay down their arms.” This caused lots of panic and confusion as the video was very believable. Because of examples like this one, researchers continue to develop methods for understanding how the video aspect of deep fakes is particularly disruptive to viewers. One study was done where researchers presented participants with fake news stories in the format of text, text with a photograph or text with a deepfake video. In their experimentation they had participants who rated the deepfake videos as convincing, dangerous, and unethical, while other participants did report false memories after viewing deepfakes. (Murphy, & Flynn, 2021)

There are technologies that are in development to help detect the danger that are deepfakes. Technologies that are being created focus on specific methods that help identify

deepfakes. These methods include; General Network-based methods, Temporal Consistency-based methods, Visual Artifacts-based methods, Camera Fingerprints-based methods, and Biologicalsignals-based methods. (Yu, Xia, Fei, & Lu, 2021.) The technologies that are being developed based on these methods are quite efficient, but not entirely strong enough to detect all deepfakes. According to Metz (2019) at this point, it is a race between deepfake production technology and deepfake detection technology on who can dominate the other first. Because deepfake technology has advanced at such a rapid rate, deepfake production has a slight edge over detection.Yu, Xia, Fei, & Lu, (2021) This ultimately shows the sense of urgency needed to speed up the deepfake detection process, and there are others who would agree with this sentiment.

With the rise of social media and technology the spread of these deepfakes can be even greater and we are already seeing this take place. Therefore more research needs to be done. The technology is advancing and we cannot sit back and wait for someone else to help combat it. Because of this advancement in technology is the motivation behind the study and ultimately thee reason why I believe that despite being technogolically savvy, college students still cannot identify the difference between real and fake.

Methodology

The methodology in my study consisted of focus groups that included anywhere between four to eight college students who currently attend St. Edward's university. The reason behind this specific selection is that college students are rumored to be the most technologically savvy

as they have grown up with the development of technology and social media and may have the best chance at identifying real videos from a deepfake.

The agreeing participants and focus groups would then join me at different areas on St. Edward's campus (e.g. dorm room, conference room, gymnasium, etc) where they signed a letter of consent and were briefed by me on what was about to occur. After getting verbal agreement and signed forms of consent, I then shared a google form in which they would be able to jot down their answers. Proceeding, I set up my laptop to allow video recording of the space to allow me to look back and observe or analyze any visual or audio responses (e.g. facial expressions, quotes, exclamatory remarks, etc.) that were to be shared during the experience.

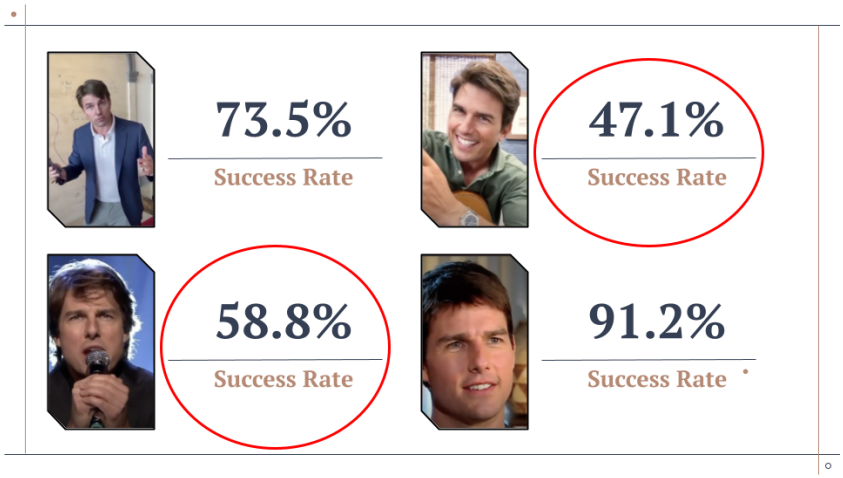
During the actual duration of the focus group, participants watched four video clips. These video clips contained four videos of famous actor Tom Cruise, where two were real clips of Tom Cruise and the other two were deepfakes of Tom Cruise. Participants would then proceed to watch all four clips strung together with audio and then watch them once more without audio to give them another perspective to analyze and then make their decision.

Subsequently, participants completed the google form answer sheet in which they shared what videos they believed to be real and which videos they believed to be a deepfake. After the form was submitted, I revealed the truth about each video, opened the floor for final comments, and then concluded the focus group session.

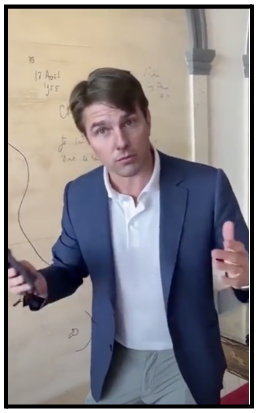
Results

Out of my c.40 participants' responses to all four videos, their total accuracy in determining which videos were real and which were deepfakes was 67%. Specifically, the success rate of identifying which videos were deepfakes was 60% and the success rate of

identifying which videos were real was 75%. However some videos were clearly harder to identify than their counterparts. Here are the results for each specific video.



Video 1:



Video 1 was a deepfake of Tom Cruise in which he trips into frame while telling a story about a conversation with Mikhail Gorbachev, Former President of the Soviet Union. This deepfake really tries to sell the exuberance that the real Tom Cruise shows in his hand gestures and body language. However it may have been slightly over done as participants had a 73.5% success rate in identifying that this video was in fact a deepfake.

Video 2:

Video 2 was a real video of Tom Cruise in which he is participating in a lipsync battle with Tonight Show host, Jimmy Fallon. In the video Tom Cruise is up close and personal as he nails every lipsynced word to the very popular pop song, Can't Feel My Face by The Weeknd. Although this video was real, users may have been fooled by the absence of Tom's real voice as participants had only a 58.8% success rate in identifying that this video was actually a real video.

Video 3:

Video 3 was the second deepfake of Tom Cruise shown in which he talks about his love for Dave Matthews, proceeds to play the guitar, and then prepares to sing one of Matthew's songs. In the deepfake Tom Cruise uses many hand gestures, but was definitely more subtle with

them in comparison to the first deepfake. This more laxed version of Tom Cruise may have helped contribute to the duping of participants as they had a study low 47.1% success rate in identifying that this video was in fact another deepfake.

Video 4:

Video 4 was the last video in the study and therefore the last real video of Tom Cruise shown. This video was an infamous interview done in 2005 on the well known show *60 Minutes,* where host, Peter Overton, is seemingly asking questions that are a little too personal about Tom Cruise and his former marriage to actress, Nicole Kidman. In the video Tom Cruise gives quick and brief answers as it is obvious he is uncomfortable and wants to move on. Truthfully this video was essentially my controlled sample in which I expected participants to get this one right and my prediction was mostly correct as participants had a success rate of 91.2%, ultimately able to distinguish that this video was real.

Analysis and Discussion

Although the participants had an overall 67% success rate, there is much more to be aware of then that specific number. My study did reveal that majority of college students can

identify a real video from a deepfake, but a third of the time participants still failed. Even further, there were specific videos that has a significantly higher failure rate than others.

For example all the videos combined resulted in an overall 67% success rate as mentioned above. However, if you only took the success rate of videos 2 and 3, the success rate dropped from 67% to almost 53%. That is a 14% decrease in success rate but also means that just about half of participants could not identify the difference between real and fake.

Prior to the execution of the study, I had expected video 1 and 4 to be easy to distinguish. Video 4 was a very popular interview and quite clear that the real Tom Cruise was speaking. In video 1, the acting is a little overdone with erratic hand gestures and exaggerated laughing. I was expecting my participants to identify those fairly accurately to which they did with video 1 being 73.5% successful and video 4 being 91.2% successful.

Videos 2 and 3 were a different story. I definitely expected video 3 to be a challenging one for the participants. The deepfake was extremely convincing as it is quite subtle in body language and the guitar playing was very believable. If there was a video I was predicting to confuse my participants it was video 3, which was definitely accurate as 47.1% (less than half) of the participants correctly identified the deepfake.

The video that was very alarming was video 2. This clip was real, but almost half of the participants thought that the video was fake as 58.8% of the participants correctly identified it as real. This was very surprising as the clip was from a Jimmy Fallon lipsync battle which is very well known, but more importantly the video was real. Going into the focus groups I felt that the deepfakes were going to skew the percentages of the success rate of the study, but the fact that a real video of Tom Cruise was one of the biggest causes for failure was shocking.

This is significant because it alludes to a world in which society can't not only identify a deepfake but also a real video. One thing also to keep in mind is that participants were aware that they may be subject to getting duped. Yet if they weren't participating in a study and rather just left to their own devices and scrolling through their social media, who knows if that success rate of identiifying a deepfake would be as high as 67%. One participant backs up this claim in which they stated during one of the focus group's concluding discussion saying, "If I was scrolling through TikTok or Instagram and saw the guitar video I wouldn't think twice." This is all extremely significant because this inability to distiuish real from fake points towards the phenomena that is Pandora's Box.

Pandora's Box is the culmination of all the potential evils and dangers in the world. If deepfake technology continues to advance and grow at such an exponential rate, Pandora's Box will open. Deepfake technology to this day has already caused many issues in our society such as blackmail, false impersonations, disinformation, etc.

Deepfakes also have the potential to alter the way we perceive reality. For example one of the most critical pieces of evidence used in our justice system is video. If videos can now be altered and manipulated that would prevent courts from being able to trust video as evidence and thus be forced to make decisions based on opinions and bias.

Unforutantely the technology to detect deepfakes is a lost cause. Each time deepfake detection makes significant advances to being effective, deepfake technology succeeds those detection services. One deepfake detection CEO stated that, "Ultimately I think it's a losing battle," Allibhai said. "The whole nature of this technology is built as an adversarial network where one tries to create a fake and the other tries to detect a fake. The core component is trying

to get machine learning to improve all the time…Ultimately it will circumvent detection tools."(Cao, 2019.)

The only real solution that could effectively combat against deepfakes is to either miliatrize it or limit deep fakes by rewriting the First Amendment in the United State's Constitution. Otherwise deepfakes and deepfake creators will always have the right to produce the artificially intelligent made content and use it for nearly any purpose they may desire.

Conclusion

Overall my study gave a glimpse of what is yet to come. Deepfake technology is already at a place to where technologically savvy users like college students can not always accurately identify real from fake. In result deepfake technology is not something to be taken lightly and it is something that ironically needs to receive more attention and regard. If we do not educate ourselves about this rapidly growing technology Pandora's Box will break open and havoc will ensue in our modern society and beyond. Until then it is important to talk about deepfakes with others, familarize oneself by watching deepfakes on their own, and find sufficient ways to distinguish a deepfake from a real video.

References:

Earnest, W., Mcglone, M., Knapp, M., & Griffin, D. (2020). Chapter 13: Visual deception. In *Lying and Deception in Human Interaction* (pp. 487–513). Kendall Hunt Publishing Company.

Đorđević, M., Milivojević, M., & Gavrovska, A. (2020). DeepFake video production and SIFT-based analysis. *Telfor Journal*, *12*(1), 22–27. https://doi.org/10.5937/telfor2001022Q

Toews, R. (2020, May 25). *Deepfakes are going to wreak havoc on society. We are not prepared.* Forbes. https://www.forbes.com/sites/robtoews/2020/05/25/deepfakes-are-going-to-wreak-havoc-on-society-we-are-not-prepared/

Murphy, G., & Flynn, E. (2021). Deepfake false memories. *Memory*, *0*(0), 1–13. https://doi.org/10.1080/09658211.2021.1919715

Metz, C. (2019, November 25). *Spot the deepfake. (It's getting harder.)*. The New York Times. https://www.nytimes.com/2019/11/24/technology/tech-companies-deepfakes.html

Cao, S. (2019, November 1). *CEO of Anti-deepfake software says his job is 'Ultimately a losing battle.'* Observer. https://observer.com/2019/11/amber-video-identify-deepfake-audio-video-shamir-allibhai/

Yu, P., Xia, Z., Fei, J., & Lu, Y. (2021, November). A survey on deepfake video detection. *IET biometrics*. https://doi.org/10.1049/bme2.12031