The Effects of Deepfakes and How We Can Mitigate Them

Prepared For
Dr. Christine Choi
CUNY New York City College of Technology

Prepared By
Alexander Marin
Cesar Suriel-Luna
Sakuna Rai

May 27, 2022

Table Of Contents

Section 1: Introduction to The Study

Introduction

Have you ever heard the story of Rana Aayub? She is an investigative journalist from India and opinion columnist with the *Washington Post*. In April 2018, while she was investigating the corruption case of the Indian government, she was attacked by the 2 minutes 20 seconds of her sex video in the internet. She couldn't believe her eyes to see such a video that she has never made. Within 48 hours of the first video released, her fake sex video went so viral that it was shared more than 40,000 over the internet. All of her online accounts were flooded with the screenshot of her nude picture, graphics filled with slurs and hatred. Even her home address and cell phone numbers were doxed and shared online along with the graphic content saying , "Available for Sex". She plunged into a deep depression that resulted in her locking herself in a room for a week and could hardly eat or speak with her family members . There was intense pressure on her from all her family members, relatives, friends and other Muslim organization questioning her, how did she dare to make such a dirty video? This is what she faced and it was all the result of "deepfake".

Deepfake is a fake video, image or audio to show people doing and saying things that they never did or said. Deepfake is generated with the help of artificial intelligence and neural networks. Although deepfake videos look real, authentic and convincing to the audience, the faces shown in the video are actually not the real one. Making deepfake for some fun things or entertainment without harming someone's personal reputation is ok, but to exploit and threaten someone's life is a serious concern. Imagine yourself as Rana Aayub and think, if you or someone from your family members are victimized in such a shameful act, what would be your reaction? Can you imagine what catastrophic result can bring in someone's life with this deepfake? The reason for choosing this topic is because deepfake can be a real threat to everyone as it can easily manipulate fake contents or news. These days, deepfakes are making it more complex and harder to trust what you see or hear. It has created confusion among the people by spreading lies and misinformation. It undermines the truth that has a great negative impact on us from political election results to personal reputations and more.

Objectives of the Research

For our research, we are seeking to analyze the misinformation caused by the use of deepfakes. Deepfakes are a fairly new technology yet we can already see so many cases in which they are used to spread lies. We will take a look at the effects that the spread of misinformation has on individuals and society. We will analyze the many instances where deepfakes were used against social media personalities and celebrities for criminal purposes. We will also look into how deepfakes are used against public figures and to control society. We'll continue our research by looking into what we can do to counter the use of this technology and prevent misinformation. We will take a look at the technologies being developed to detect videos that are deepfakes. Furthermore, we will look into what legislation can do to curb the use of this technology and the difficulties behind doing so.

Methodology

For our research, we did different kinds of searches about deepfake in the City Tech library. For example, we use combinations of terms such as detection, entertainment, misinformation, and politics with the term deepfake. With these searches, we ended up finding a lot of articles but just a few educational books on deepfake. Since deepfake is considered to be still new to the world and developing at a rapid pace, this was the reason we ended up with mainly articles as sources. With a lot of articles, we had to narrow it down to peer-reviewed articles. In addition to the City Tech library, we used the google search engine to get different sources about the situation of deepfakes. We wanted sources that mention the actual misinformation that deepfake is spreading and how it is affecting society as a whole. Our research also involved looking for a deepfake video that we can use as an example for those who quite don't know what deepfake is exactly.

Section 2: Research and Results

How the Technology works

There are two different types of deepfake technology which are face swapping and facial reenactment. Face swapping is a method that can replace the face in an image with the same facial shape and features as the face being implemented. Facial reenactment is a method that can manipulate certain facial attributes and reenact reactions with deep learning methods. There are also detection methods according to the article "Deepfake generation and detection, a survey," One of the methods is biometric features such as eye blinking, lip-sync facial and head movements, head pose, color, texture, and shape cues. These are human features to look out for in deepfake videos due to the difficulty of replicating these features. There are also model features, and media features to detect deepfake technology. Model features look for specific model fingerprints that can be left behind when developing a deepfake video, but the creator of a deepfake content can remove these fingerprints successfully. Media features are used to detect temporal information, the inconsistency between frames in the deepfake video, and noise artifacts.

Generally, Deepfake is created by artificial neural networks, which means a computer learning to perform special tasks by doing the same task repetitively. According to the article, "Deepfakes, Real Consequences: Crafting Legislation To Combat Threats Posed By Deepfake", the neural network is trained to automatically map the facial expressions of the source to create the manipulated video. (Janga, 2021, 764). In a single neural network, there is no any detector or tool to confirm how convincing your fake video is. Nowadays, deepfake videos are made using two neural networks trained to work in tandem which we call GANs (generative adversarial networks) in short. One neural network is called "actor" which tries to copy all the moves and expressions of the original video or audio into fake videos while another neural network is known as a critic task is to distinguish between the real and fake contents trying to outwit each other until the critic can no longer tell which is real and fake.

Currently in the world right now, the impact of deepfake technology was unforeseen by intellectual property policymakers at the time current laws were created. Current laws helped

performers be legally entitled to their records which are the performer's rights (property rights), but not a digital impersonation of themselves. The article, "Giving performers copyright over their work could protect them from deepfake technology, study shows", explains reforming performers' rights is the best way to deal with the challenges that deepfake brings. Establishing the performer's rights to copyright can help performers in a big way and be protected from unauthorized deepfakes. Deepfake technology can be used positively for creativity, or entertainment industry such as television, films, video games, and social media. All of this means that performers can use the positive aspects of deepfake technology, and not let the public be confused about whether or not it's their actual creativity/performance or not.

Deepfakes and Sexual Abuse

Deepfake first started in pornography on the platform Reddit where an anonymous user posted videos in 2017 that featured famous actresses' faces on other women's bodies. This then created worries with national security experts. The article, "Going Deep: In a world of 'fake news' accusations, deepfakes may soon be a very real problem for journalists' ' describes it as a dystopian future. A future where society would begin to question if any video or audio can be trusted, and where deepfake can be used for blackmail, ruin a victim's reputation, and serious damage. Serious damage such as riots or even violence can be caused by misinformation by deepfakes. Luckily, according to the article, there is a media forensics committee where journalists can undergo specialized training and workshops about deepfakes and the tools and methods to detect them. They also allow journalists to contact the committee for help to see if the video is credible.

Since Deepfakes can have numerous immediate effects on social media platforms, media outlets and governments, they can cause or bring huge catastrophes in personal life, family, society and nations. For instance, as mentioned earlier in the introduction section, how Rana Ayub was being attacked with a porn plot. In the article, "*I Was The Victim Of A Deepfake Porn Plot Intended To Silence Me"* written by Ayub, herself is a shocking story. She had explained all her embarrassing and frustrating moments as being a victim of deepfakes that she had never known. She was admitted to the hospital with anxiety and heart palpitation problems. Doctors gave her medicine but she was vomiting because her body was reacting violently to the stress. Eventually, the United Nation had to appeal to the Indian Government to protect her. So, we see in the article how dangerous is a deepfakes and its immediate effect on someone's life.

The Spread of Misinformation

Deepfake technology has been grossly misused in a few different ways, especially for spreading misinformation. Videos get posted on social media for the world to view and the truth

is many people view things at face value. A video of someone saying something offensive will get posted, and rather than doing any further research, to some people what a political candidate or public figure says becomes fact. For my research, I conducted a small study to determine just how many people would believe a deepfake video without questioning it. For our study, we used a video created by artists Bill Posters and Daniel Howe which shows a deep faked Mark Zuckerberg speaking bluntly about how Facebook's mission is to profit off of the data of its users. Of course, Zuckerberg has never made such comments, but we wanted to see how many people would believe the video. We created a google form where participants would watch the video and give their thoughts. After writing in their thoughts the next page reveals that the video is fake and asks if they realized that the video was a deepfake, and if they did then what gave it away. The results of the survey showed that 30% of students questioned believed that the video was real. Some responses we got said “I feel that now I don’t feel like using Facebook because they basically not giving me privacy how he telling me that he owned me” and “I think this video is telling the truth because there’s no such thing as privacy on the web especially social media”. It can be seen that videos like this can cause people to form some strong opinions. The survey did also show that deepfake technology isn’t quite powerful enough yet deceives most people and there were faults in the video that made it obviously fake. When we asked “Did you feel like there was anything strange about the video while you were watching it?” We got responses saying things such as “How the person was speaking and moving their lips it didn’t seem normal” and “It looks like a deep faked based on the lack of facial expressions” .

Deepfake technology can have some glaring issues behind it although it still hasn’t stopped attempts for it’s use as a weapon against people. In Bobby Allyn’s article “Deepfake video of Zelenskyy could be 'tip of the iceberg' in infowar, experts warn” on NPR, we get some insight on how deepfakes are currently being used in warfare and the damaging potential they can have in the future. The article explains that in the midst of a war with Russia, a deepfake video of Ukrainian president Vladymyr Zalensky telling his soldiers and civilians to lay down their arms and surrender. The video, which is believed to have been created by Russia, was spread around social media. Through hacking, the video was broadcasted on Ukrainian television, and posted on the website of several news outlets. This was one of Russia's many attempts to demoralize the people of Ukraine through information warfare. The video however, was quickly debunked by the state and proven to be fake. The article says that many people noticed that Zalensky’s accent was “off” and that his head seemed too large for his body. However, the article also notes that while Ukraine was able to get ahead of this video weeks in advance by warning its citizens about the possibility of such videos, the video can have other effects. In the article, when Bobby questions Hany Farid, a professor at the University of California, Berkeley who is an expert in digital media forensics, he states “It pollutes the information ecosystem, and it casts a shadow on all content, which is already dealing with the complex fog of war…The next time the president goes on television, some people might think, 'Wait a minute — is this real?' " When you put this situation in perspective you can realize how

much damage that even a debunked deepfake can have. They can have a lingering effect where people will approach everything that they see with skepticism.

Nowadays, there are many fears about deepfakes being used in politics. In a journal by Andrew Ray published by the University of New South Wales Law Journal, he researches and discusses the effects that deepfakes have had on politicians and their campaigns around the globe. One case Ray discusses is of Belgian prime minister Sophie Wilmès where during her election campaign in 2020, a deepfake video of her talking about the link between COVID-19 and global. As with previously discussed cases, the video was proven to be fictitious although some people still believed the video. Ray mentions how deepfakes have circulated of many other politicians such as Donald Trump, Barack Obama, Nancy Pelosi, and Joe Biden. While many deepfakes that are made are for the sake of comedy, a few are also made to discredit these people and seem out of touch with their base. Ray says that this can have grave effects on these politicians, their parties and during elections where even a few misled voters can make or break a victory for either side. Ray takes note of how different digital consumption is compared to just a few years ago and the effect it has had on the public perception of politics. He says "Increasingly, Australians are turning to digital platforms such as Facebook to access news content... The shift to digital content has coincided with decreasing trust in politicians and politics in general. Political deepfakes will further erode trust by allowing candidates to deride real footage as fake news, feeding into increasing claims by politicians that they have been set up". The adoption of social media networks as news sources has resulted in increasing political distrust. The development of deepfakes can result in further damage if politicians become able to wave off real videos as fake if people don't know what to believe anymore.

Detection

Researchers at the Albany University have discovered one such tool to detect deepfakes is examining the eye blinking rate. According to the article, *In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking.* "We focus on the detection of the lack of eye blinking to expose AI synthesized face videos." (Li, -et. Al., 2018). In their research , the team was able to expose the differences in the eye blinking of deepfakes compared with authentic sources. According to the article, the suspected images or videos are passed through diagnosis tools like Long-term Recurrent Convolutional Networks or Convolutional Neural Networks. Both LRCN and CNN are artificial intelligence neural network applications, created based on human neuron concepts. We can scan the whole images or face of the suspected sources in LRCN and CNN, but this will consume more time and maximum memory allocations are required for each scanning process. So it is not a good idea to scan the whole image or video, instead we have to scan particular objects like eyes, nose, lip, cheek or forehead, etc. separately to get the result done more efficiently and faster. But the problem is, not all objects like cheek, forehead, nose, etc. give us a more convincing result because those objects hardly change their skin tone and behaviors in each frame of the video. Thus, the eye blinking is considered to be

the most convincing object to scan the deepfakes, says the article. Such eye blinking is categorized into three major types - spontaneous blink, reflex blink and voluntary blink. Each eye blinking has a different blink rate and blinking duration which further helps in comparing the deepfakes with legitimate sources. While this approach of exposing deepfakes shows significant promises, there are still numerous challenges unexplored. What if someone creates realistic eye blinking effects?

Discussion

From looking at the research we have conducted on deepfakes, it's clear that its ability to be used maliciously is cause for concern. We look at the results by firstly looking at who has been affected by deepfake technology and seeing the consequences of it. We can see that deepfakes can affect a wide range of people. Many women become victims to it through having their faces swapped onto pornographic videos. Videos such as these are damaging to ones representation and can be considered as grounds for sexual assault. Politicians can have their political campaigns damaged by the creation of false statements. Deepfakes can even affect a whole nation when used for the purpose of information warfare.

We have seen so far, how deepfake had become more popular in 2017, after the celebrity's faces were swapped to porn stars. Then this trend of swapping faces continues to grow significantly resulting in threatening or harming someone's reputation. As a result of that, today we are facing a greater threat of misinformation posed by deepfakes than ever before. We all are facing the same fate of consequences caused by deepfake rather than its unseen advantages. Another surprising thing is right now there is no ultimate solution either to detect deepfake or any concrete law to combat the threat posed by deepfakes. In some countries like China, where deepfakes are totally banned, there are no issues at all, but countries like the USA, India, etc. where only partial laws related to deepfakes are drafted have to face many serious challenges like national security threat to personal threat. As a result of this there are many victims like Rana Ayyub, to whom the United Nations had to appeal for her protection.

The research showed that there was no proper way of protecting someone from it and had no serious consequences for the people who created the deepfake video in the first place. All of it which was proven by the story Arub wrote about what she went through. Her story also proved her reputation was ruined and couldn't live a normal life. The results also tell the backstory of deepfake and how it began in the industry of pornography first rather than the film industry. Once again was proven by Arub since she was the victim of a deepfake pornography video, What the results also show is that there are methods of detecting a deepfake video that could be a solution to solving whether a video is deepfake or not.

Moreover, all the journalists from news readers to investigative journalists, who themselves do intensive research on various topics have to undergo specialized training and workshops to recognize the tools and methods to detect deepfakes. For them, deepfakes have become the most hated topic to be discussed today and hard to digest in their professionalism. We have seen that Ayyub, who herself is an investigative journalist and columnist could not

protect herself from being attacked by deepfakes, how could one civilian protect him/her from being attacked by a heinous act of deepfakes?

While it can be argued that deepfake technology is not advanced enough to be very convincing in many cases, the effect it has on even a few people can play a larger part in affecting a group. The use of deepfakes in this digital information age can lead to high levels of distrust for anything that we see online. Now, through the increasing use of social media the effects are becoming increasingly amplified. Deepfake technology is still being developed and as AI and machine learning improves, so will the quality of deepfakes.

Section 3: Recommendations and Conclusion

Conclusion

To conclude, the recent advancements of deepfake technology have made it an important matter of discussion. When diving into the effects it has had on our society the problems it causes are nothing short of apparent. Deepfake technology is not that simple and easy to fight against if used wrongly. Various research institutes and government agencies are intensively working to tackle the rising issue of deepfakes. As mentioned earlier, it has affected quite a number of people already and the research has shown anyone can be a target of deepfakes, from individuals, to an entire society. The problems raised by deepfakes have been pointed out by many scholars but solutions for solving these issues are still under question. So, while the issues are apparent the solutions are not. Currently we don't have any laws that are made to directly target the use of deepfakes, in fact the laws we have now protect them. To add on, while machine learning seems promising, deepfake technology is also advancing rapidly and could develop to the point where it may be undetectable even to other machines. The best measure that we can take right now to curb this issue is informing the people. Even though there are some highly technical methods to detect deepfakes such as AI detections and biometrics, it will be up to the public and media to determine whether or not a video is authentic or not. Poor quality deepfakes are easy to spot by paying close attention to the quality of the deepfake. For instance, the patchy skin tone, the agedness of the skin mismatching to the agedness of hair and eyes, lip syncing failing to match with the size and color of the lips and other objects like facial moles and dimples looking unreal. But not all the deepfakes are always poor in quality and it is not an easy task for the human brain to spot those patchy skin tones, lip sync or mismatching agedness of skin to the agedness of hair and eyes. So, we have to come up with a different approach to identify those deepfakes which is only possible with either a strict law or advanced artificial intelligence itself. Moreover, it is also the responsibility of an audience to be aware of where the source is coming from. People need to be educated on what it means to have a reliable source and the signs of a deepfake video.

Recommendations

When discussing the solutions for the problems caused by deepfakes there are many things to consider, the first thing we must consider is that there is no clear solution. No matter what we do, the moment a video gets posted it can already be too late. The main things that need to be focused on are curbing the posting of these videos, creating legislation against such videos, and educating people on the importance of fact checking what they see.

As iron sharpens iron, so is with the deepfakes. As deepfakes are created with artificial intelligence, the best possible way to detect deepfakes is, the use of advanced artificial intelligence itself. According to an article called "Deepfake Detection: A Systematic Literature Review" published by the IEEE, there are several models that are proven to be able to detect deepfakes including deep learning models, machine learning models, and statistical models. The best of these models is the deep machine learning convolutional neural network (CNN) model which according to the report was able to detect deepfake videos made in a controlled environment 78% of the time. Technology like this can be implemented into media websites that can detect and flag these videos before they become posted. They can even track the source of the origin of suspicious contents (images or videos) where it's coming from and can block content before it's uploaded to their site. Taking this into consideration, we can use a system like this for social media and news outlets. Youtube, for example, has already implemented a system where it restricts certain videos that have copyrighted music, or even has immature content in the video. Social media institutions like Youtube, which is where most information is spread, can have a deepfake detection system before a video goes online to the public. This is similar to the idea given by *NewsRx Science* in their article "Giving performers copyright over their work could protect them from deepfake technology, study shows." where they believe copyright should be expanded more to give performers protection from deepfake technology.

While the creation of legislation has been proposed there are many challenges in doing so. In Andrew Ray's journal, he discusses how laws surrounding copyright can be used to protect individuals. He mentions how the Kardashians were able to remove a deepfake video off of Youtube using YouTube's copyright infringement procedures although the video was still able to be seen on other platforms such as instagram. However, Ray also mentions how copyright fair use laws can also protect deepfake videos as the Copyright Act provides exemptions to works that were created for parody or satirical use. We have laws, but there is no specific law addressing the whole issue of deepfakes in detail. Besides this, each state and country has their own law. We cannot regulate the laws across the world. The deepfakes banned in one state can be available in another state or deepfakes banned in the USA can be easily available in other countries like India or France, etc. We can have control of deepfake over one country, not globally.

To combat deepfakes, they need to be well known by the public and the media. Once deepfake technology gets well awareness by the public, the public would question the authenticity of a video and research whether it's true or not. Public education about deepfakes can end up being one of the most important tools for stopping the spread of misinformation. As we discussed with Ukraine, states telling their citizens about the possibility of such videos help with the quick debunking of deepfakes when they arise. Educating the people on sources they can trust is paramount for keeping misinformation under control.

References

Allyn, B. (2022, March 17). Deepfake video of Zelenskyy could be 'tip of the iceberg' in Info War, experts warn. NPR., from https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia

Andrew Ray. (2021). Disinformation, deepfakes and democracies: The need for legislative\ reform. University of New South Wales Law Journal, 44(3), 983–1013

Ayyub, Rana. I Was The Victim Of A Deepfake Porn Plot Intended To Silence Me. (2018, November, 18) https://www.huffingtonpost.co.uk/entry/deepfake-porn_uk_5bf2c126e4b0f32bd58ba316

Giving performers copyright over their work could protect them from deepfake technology, study shows. (2021, September 26). *NewsRx Science*, 164. https://link.gale.com/apps/doc/A676046139/AONE?u=cuny_nytc&sid=bookmark-AONE&xid=d131a8ec

Langa, J. (2021). Deepfakes, Real Consequences: Crafting Legislation to Combat Threats Posed by Deepfakes. *Boston University Law Review*, *101*(2), 761–801. http://citytech.ezproxy.cuny.edu:2048/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=150097538&site=ehost-live&scope=site

Li. Yuenzun, Chang, Ming-Ching, Lyu, Siwei. (2018, June,11). *In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking.* https://arxiv.org/pdf/1806.02877.pdf

Md Shohel Rana, Mohammad Nur Nobi, Beddhu Murali, & Andrew H. Sung. (2022). Deepfake Detection: A Systematic Literature Review. IEEE Access, 10, 25494–25513. https://doi.org/10.1109/ACCESS.2022.3154404

MORRIS, A. (2019). Going Deep: In a world of “fake news” accusations, deepfakes may soon be a very real problem for journalists. Quill, 107(2), 21–25.

Zhang. (2022). Deepfake generation and detection, a survey. *Multimedia Tools and Applications*, *81*(5), 6259–6276. https://doi.org/10.1007/s11042-021-11733-y