10-2023

Implications of Deepfake Technology on Individual Privacy and Security

Shalini Mahashreshty Vishweshwar

Implications of Deepfake Technology on Individual Privacy and Security

By

Shalini Mahashreshty Vishweshwar

A Starred Paper

Submitted to the Graduate Faculty of

St. Cloud State University

in Partial Fulfillment of the Requirements

for the Degree

Master of Science in

Information Assurance

December, 2023

Starred Paper Committee:
Jim Chen, Chairperson
Lynn Collen
Balasubramanian Kasi

Abstract

Technological advancements not only can make life easier but also create menacing consequences when misused. Deepfake technology, being one of the major advancements recently, serves as an example of such technology. It is extremely difficult to identify the fake media from the real media. Deepfake technology uses Artificial Intelligence to impersonate someone and create hyper-realistic media like videos and pictures. This paper assesses the rising implications of Deepfake technology on individual privacy and security. The augmentation of deepfake capabilities, powered by Artificial Intelligence, gives rise to a predicament in accurately discerning authentic media from concocted variations, thus jeopardizing personal identities and data integrity. The study examines the emergence of Deepfake technology, exploring its societal and commercial ramifications. It confronts the adequacy of prevalent legal frameworks and the urgency for upcoming regulatory advancements. Moreover, the research delves into the range of countermeasures, from detection techniques to instructional initiatives, designed to restrain the malignant exploitation of deepfakes. The emerging discourse unearths a burning demand for an integrated approach that includes legislative action, technological safeguards, and heightened public awareness to straddle the dangers brought by this double-edged technology.

Acknowledgements

I would like to express my sincere appreciation to all those who have contributed to the completion of this study. First and foremost, I would like to thank the Chairperson of the Committee, whose insightful comments and suggestions have helped to shape this paper into its final form.

I would also like to acknowledge the valuable guidance provided by my supervisor and the committee members whose support and encouragement has been instrumental in our research process. In addition, I extend my gratitude to the various individuals and organizations who have generously provided their time, resources, and expertise during our research.

Without their contributions, this research would not have been possible. I hope that this paper will contribute to a better understanding of the implications of deepfake technology on privacy and security in society, and I look forward to continued discussions and advancements in this field.

Table of Contents

List of Figures

Figure **Page**

Chapter I: Introduction

Introduction

Information technology has improved tremendously over the last decade. Unfortunately, misuses of technology are also on the rise. One such misuse is deepfake media which has been negatively impacting society and their discourse.

Deepfake is an artificial intelligence technology, which can create hyper-realistic media such as images and video. This made it possible to create audio or video of a real person saying and doing things he or she never said or did. Deepfake can be highly deceiving and dangerous as it has a high potential to manipulate the public's opinions and their decision making. It may also create havoc in victims' lives. While the deepfake initially targeted on political leaders, celebrities, and artists, it may extend its misuse among ordinary people. For example, it can be used in creating porn videos as a bullying, revenge, and extortion tool. As technological advancement is inevitable and so are their threats to people if misused, it is highly essential to become aware of such technology and to create a proper plan to address the issues involved with it.

This paper presents an overview of the emergence of Deepfake technologies, their benefits and usage, means of detection, societal impact, and measures to mitigate misuse. It delves into existing legislation, regulations, and policies aimed at regulating such incidents. It discusses training and education on deepfake technology and recent technological developments to detect deepfake media. The remainder of this chapter includes sections on problem statements, the nature and significance of the study, limitations, and definitions of some key terms.

Problem Statements

1. Public Understanding: There is a significant lack of public understanding regarding deepfake technology, its capabilities, and its potential risks. This gap in knowledge hampers effective public action and decision-making.
2. Creation and Detection of Deepfake: Currently, there exists a noticeable gap in comprehensive research concerning evolutionary trajectory of deepfake and concurrent advancements in detection techniques. This absence of an in-depth understanding hinders the development of robust countermeasures to mitigate the potential misuse of Deepfakes.
3. Social Impact: There is a significant gap in comprehensive analysis addressing the broader societal, ethical, and legal challenges arising from the proliferation of Deepfakes. This gap hinders the formulation of informed policies and guidelines to navigate and mitigate potential adverse effects in these critical domains.
4. Legislative Actions: There is a scarcity of research capturing public opinions on governmental legislative actions aimed at reducing the risk of deepfake technology. Timely public opinions can help lawmakers in the creation of policies that are both effective and publicly supported.

Nature and Significance of the Problem

There is a great need for people to understand the risk of deepfake technology. By addressing the gap in public understanding of deepfake technology, the government and educational institutions can make better policy decisions. This is important for democratic societies where public opinion can influence policy decisions. Understanding the public's awareness level can also help technology companies to develop effective deepfake detection tools, making technology a part of the solution. By understanding and addressing the potential harms of deepfake technology, the study contributes to maintaining social trust and it can be eroded by malicious uses of the technology. Finally, the study can provide a basis for educational programs and training materials to counter the impact of potential misuse.

Study Questions

- What is deepfake, and how did it emerge?
- Where is deepfake technology currently used?
- What are the benefits and disadvantages of deepfake?
- What are potential harms caused by misuse of deepfake technology?
- To what degree is society aware of the Deepfakes?
- What are the mitigative measures available currently to alleviate the damage caused by misuse of deepfakes?
- What are the available regulations to curb the misuse of deepfakes?

Limitations of the Study

As a novel technology, the literature on the impacts of deepfakes is limited, and empirical research on its psychological effects is scarce. Furthermore, it was challenging to find specific cases of deepfake misuse due to confidentiality and legal restrictions, particularly in cases of deepfake pornography, where actual incidents may not be accurately reported due to the defamatory nature.

Definition of Terms

Deepfake: A video or an image of people in which either their face or body are digitally modified to make them appear as someone else.

Fake news: It is a type of journalism which runs on with an aim to deliberately spread misinformation and false media on news platforms or social media.

Artificial intelligence: It is the ability of a computer to do tasks that are usually done by human beings where human intelligence is required.

Supercharging scams: The scams in which deepfake audio is used to impersonate the person on the other line is a higher-up such as a CEO and soliciting an employee to send money.

GPT-2: Generative Pretrained Transformer-2- is an AI model which can predict the token in the upcoming sequence in an unsupervised way. GPT-2 is a text generating AI released by the research lab named OpenAI.

GAN: GAN stands for Generative Adversarial Networks which is one of the latest advances in the deep learning technologies which deals mainly with image recognition, data computation, and broader analysis which involves activities like assessing and recapitulating the main four macro-environmental factors such as political, economic, socio demographic and technological that are the basis of the main changes that take place in the world. (CVISIONLAB, n.d.)

RNN: Recurrent neural network - type of artificial neural network which uses sequential data or time series data. Long Short-Term Memory is a deep learning architecture used in RNN.

Markov Chain: Markov Chain is a mathematical system that experiences transitions from one state to another according to certain probabilistic rules.

Chapter II: Background and Review of Literature

Introduction

In this chapter, we provide background information on deepfake technology, discuss related incidents, conduct a literature review addressing the problem, and outline the methodologies used to combat these issues. The literature review includes statistics and figures illustrating the damage caused by the misuse of deepfake content. Additionally, we delineate the methodologies based on existing works on deepfake technology.

Background Related to the Problem

Deepfake technology is developed using artificial images and audio files that are consolidated along with machine-learning algorithms with an intention to manipulate the media to create false information. Deepfakes can cause high risks, such as undermining cybersecurity and influencing political elections. They can also impact the financial conditions of corporations and individuals, tarnish the reputation of individuals, organizations, and communities, and lead to disruptions in the lives of individuals in several ways.

Initially, deepfake's focus was only on celebrities. However, now-a-days, the deepfake technology is easily available to the ordinary people and they are able to develop their own deepfake content and due to this, issues arise such as trying to manipulate the society and public, invasion of personal space and can attack rights of individuals, extortion scams, financial frauds and Supercharging scams.

As reported by a visual threat intelligence company named Sensity Systems Inc, (Petkauskas, 2021), the various deepfake videos that are generated by the AI powered

GAN (Generative adversarial Network) have shown overwhelming rise in the reputation attacks. Not long ago, a report named ‘The State of Deepfakes 2020’ had published that more than 85 thousand subversive deepfake videos have been created by the expert designers which were tracked down until December 2020. Since the year 2018, it has been found that the volume of deepfake videos that are generated by the expert crafters has doubled for every six months and these deepfake videos were certified as the videos that were used to either cause harm to the luminaries or potential enough to do so and this report excluded the list of videos that perform attacks on the individuals.

As mentioned by Giorgio Patrini, CEO, and cofounder of Sensity, there is currently a great expansion of existing communities of deepfake technology developers and corresponding content creators. Simultaneously, novel deepfake communities are emerging globally. Patrini, mentioned in one of his interviews with CyberNews that pornographic content and deprecatory, derogatory videos that cause reputation attacks and character assassination, shares a major part of 93% in the list of existing deepfake videos and also it is the West of the United States that has been a major target in terms of the misusing technology in attacking Celebrities.

Apart from the reputation attacks on celebrities, the ordinary people are also facing issues of being targeted by the personal attacks performed using deepfakes. In one of the reports published in Fall 2020, by Sensity, it is found that a bot network on the Telegram platform is developed to take photos of women from their social media accounts and manipulated to stripped off clothing using AI technology which if observed shows how much havoc it can cause in their daily lives. It is found that over 100,000 women were victimized by male offenders.

One of the best live examples on how deepfake can cause tremors in the lives of people globally is the deepfake video that had been created on then President of USA, Donald Trump talking offensively on Belgium’s climate policy (Burchard, 2018). This video was created by one of the local Political Party in Belgium named Socialistische Partij Anders shortly known as Sp.a and they posted it on twitter and Facebook. It has created a great aggravation globally and provoked the people to add hundreds of comments voicing their outrage on then American President that he would dare to interfere with the Belgium’s Climate Policy and had hate comments on American culture as well.

However, later the SPA political party confessed that it was done by a team commissioned by the political party to use machine learning to produce a deepfake video and posted it with an intention to initiate a public debate to attract attention to the climate change act. They also claimed that the video was not intended to deceive supporters but to bring attention to the issue of climate change and said they checked it is legally acceptable to do so.

Although, the Belgian political party claimed that it is legally acceptable to deliver a video campaign with the deepfake video, it has created considerable amount of disturbance across the countries and triggered an unnecessary chaos among public and increased hate on Americans which is not acceptable. However, this situation shows that deepfake is not enough to materialize as a threat to democracy and this shows there is a need to incorporate stringent rules on usage of deepfake technology in the upcoming future.

As mentioned earlier, deepfake is a phenomenon which is a combination of Deep Learning technology and fake media that are developed using artificial intelligence technology. It is found that it takes just 300 images of certain person who is a victim of deepfake to develop a reasonably convincing deepfake media using a swap technique in which faces of the source and the target are swapped effectively so that it seems hyper-realistic. Irrespective of the technic being employed for the creation of deepfake the process comprises of three basic steps such as extraction, training and creation which makes it much easier even without having huge data and a single image of a source will be enough to create a deepfake in near future. Having said that, there exists various issues related to deepfakes, Betül Çolak has emphasized legal issues like Intellectual Property rights and personal data protection in the article “Legal Issues of Deepfakes.” (Çolak, 2021).

He reported that, according to WIPO (World Intellectual Property Organization) as published in “Draft Issues Paper on Intellectual Property Policy and Artificial Intelligence” essentially two questions addressed in it.

1) As the deepfakes are created using the content, subjected to copyrights, who should be given the copyrights of the deepfake? The source or the creator?

2) Is there a necessity to have a system of equitable remuneration for persons whose likeliness and “performances” which are used in deepfake? (WIPO, 2019).

WIPO indicates that the deepfake content has immense potential to cause severe issues such as infringement of human rights, privacy rights, personal data protection right and so on and so forth. Hence, the WIPO claims that the main concern should be if the copyrights to be even consented to the deepfake media rather than the concern of whom should be given the copyrights. In response to this concern, WIPO asserts that if the deepfake content significantly diverges from the victim's identity, it should not be granted copyright protection. In cases where copyright applies, it should be attributed to the creator of the deepfake, given that there is no involvement of the source person whose image or other media is used in the creation of the deepfake, and it is done with their consent. It also indicates that copyright is not an appropriate weapon as the victim of deepfakes do not possess an interest in copyright of their own image. However, the victim can claim the right of personal data protection to overcome this issue of unethical use of deepfake.

According to the Betül Çolak, a lawyer specializing in IP and Technology law from Turkey and a Researcher in the AI and Fairness research cycle of the Institute, both the states in USA and the BigTech is actively involved in taking action against the issues of the deepfake content. A best example of this is the active development of the detection tools for identifying the challenges faced by the deepfake content by the BigTechs and the introduction of regulations in the states of Virginia, Texas and California, which are one of the first states to do that in which the law states that there exists criminal penalties on the circulation of the non-consenting deepfake pornography in the state of Virginia whereas in the state of Texas, the law forbids the people to create and distribute the deepfake media that contemplates to inflict upon the candidates of public office or intended to manipulate the election results. Considering the current issues pertaining to deepfakes, there is a great need for developing the technological tools and stringent regulations against deepfake in order to prevent as well as cease the destructive outcomes of the misuse of deepfakes.

Literature Related to the Problem

According to a report on effectiveness of deepfakes in manipulating the attitudes and intentions of the people generated by conducting experiments on people by exposing them to the genuine and deepfake media and measured their explicit (self-reported) and implicit (unintentional) attitudes as well as behavioral intentions and the results obtained from this experiment indicates that the deepfake media such as video, audio and images have a severe psychological impact on the audience i.e., viewers and it is just as effective as the genuine content has in manipulating their attitudes and intentions. (Hughes et al., 2021). In their study, they have conducted seven preregistered experiments with an aim to find out what can happen if a viewer happens to come across deepfake content and what will be the impact of being exposed to deepfake content just for once and how it affects their biased thinking compared to the actual legit content.

The results suggested that the deepfake video content creators can easily manipulate the attitude and thinking of the people who are exposed to it thereby making them vulnerable to giving up control to these deepfake content creators which is a kind of violation of personal right to think and act rightly as the information that is being put forward is fake and hence the actions would also be wrong. This study highlighted the most dangerous aspect of deepfakes, their capability to undermine our beliefs in what is reality and what is the reliable information that we can trust. 'Liar’s Dividend' is a concept that suggests that some people misuse deepfake technology with the intention of profiting from the data environment, which is deluged with fake information.". (Chesney & Citron, 2018)

Deepfake technology had originated initially in the computer vision field and later, subsequently found its usage in audio manipulation and then the text generation (Bregler et al., 1997). In the recent times, there has been lot of advancements in software which are capable enough to generate the speech audio text body by manipulating the voice of various speakers just after listening for five seconds (Jia et al., 2018). According to a group of deepfake text detection is highly essential as there exists development of exceptionally advanced generative methods such as GPT-2, RNN, LSTM, Markov chain. However, it is found in their study that there are not enough methods or software available to detect the deepfake texts that are visible in huge numbers on social media texts yet. Also, their study found the existence of 25,572 tweets, comprised of half human and half bots crafted and posted on twitter. (Fagni et al., 2021).

According to Fred Eslami (Zeman, 2021), an associate at AM Best, a credit rating agency, says in a world of insurers, the cyber incidents are unique in nature and there is not much steadfast historical or factual data that can be utilized to predict the depredations in cyberattacks unlike the natural disasters where the factual and historical data exists and can be used to predict and prevent the losses. Unlike the regular cyber-attacks, deepfakes use artificial intelligence to misrepresent the recorded audio and video content. Although, initially deepfakes target was movies and entertainment purposes like humor, eventually deepfakes were used in changing the factual data and spreading the misinformation which poses the deepfake content a highly threatening tool and the social media will aid in making them even more dangerous as it is the medium which spreads the deepfake content instantly across the globe.

According to Cybercube, a Saas Company, cyberattacks are increasing at a high rate using social engineering and various other cyber procedures and when the deepfake technology meets these techniques it will only raise the success rate of these attacks exponentially high (Zeman, 2021). One such incident took place in March 2019, where the cybercriminals used AI based software to mimic the voice of Chief Executive Officer of a UK based energy firm with an intention to impersonate him to gain unethical profit of fraudulent transfer of the €220,000 ($243,000) which was identified by cyber-analysts as an unusual case of AI usage in hacking activities. The cybercriminals were successful in their attempt to transfer amount to a Hungarian bank account and then distributed to other locations and due to this the investigators could not trace back the hackers and it caused a huge loss the company which is extremely unfair to the victims. Many times, cyber-attacks cannot be traced easily and may not even recover the losses in these cases. (Stupp, 2019).

Deepfakes and the Law: Current Legal Approaches and Regulations on Deepfakes

Deepfakes have become a prominent issue today due to their potential to deceive and manipulate individuals through the synthetic alteration of audiovisual content using AI technologies. These sophisticated digital fabrications can blur the boundaries between truth and falsehood, presenting significant challenges across multiple domains such as politics, privacy, security, and trust. Recognizing the potential risks and harms associated with deepfakes, policymakers and legal experts have been engaged in a complex and evolving discourse regarding the development of effective laws and regulations to tackle this emerging technology.

At present, there is a lack of comprehensive legislation worldwide, specifically targeting the issue of deepfakes. Each country has its own set of laws and regulations, leading to variations in approaches. However, it is evident that most countries have not established dedicated legislation specifically aimed at combating deepfakes. Instead, existing privacy laws and data protection regulations are often utilized to address instances of deepfake misuse.

According to Scott Briscoe, a Content Development Director at ASIS International in his article on Laws addressing the deepfakes, there are stipulations included in U.S. National Defense Authorization Act (NDAA) to address the growing problem of deepfakes recently (Briscoe, 2023). According to 2021 NDAA, a law has been created to issue and annual report on deepfakes by Department of Homeland Security and it should encompass a comprehensive examination of potential risks and damages associated with the technology, as well as addressing a wide range of concerns like foreign influence campaigns, fraudulent activities and harm inflicted upon specific group of people. This has aided in broadening the scope of deepfake report mandate that was called previously by NDAA.

According to a report published by Hogan Lovells (Lovells, 2020), a Global law firm, the European countries like UK, France or Germany has no explicit legislation in place that directly addresses the legal framework pertaining to deepfakes. In case of any online disinformation including misuse of deepfakes, the European mandate intends to address these issues through a set of measures like self-regulatory Code of Practice on Disinformation for online platforms. This code aims to ensure these online services have security measures against disinformation and checks on availability of appropriate tools to report the disinformation.

Currently as there is no legislature to combat the misuse of deepfakes present in European countries, they utilize existing laws, such as laws on deepfakes and the right to protection from derogatory treatment, to address the deepfake situation as a workaround.

In case of United states, according to Hogan Lovells report, various states have recently enacted laws aimed at addressing the detrimental effects of deepfakes. However, these laws undergo substantial scrutiny by the First Amendment rights of free speech, and it is uncertain whether courts will determine if these state-level regulations violate Constitutional principles. Some of the laws that have been created to combat misuse of deepfakes as listed below:

- Couple of California laws made effective in 2020 which regulated the distribution of deepfakes including altered images, audio, or any visual deceptions of a reasonable individual (Tremaine, 2019). The two laws are described as
    - AB 730 – restricting the usage of deepfakes in Political campaigns manipulation and it comes with an expiry date on January 1, 2023.
    - AB 602 – aims to tackle the issues regarding deepfakes and pornography and do not have any expiration date.
- Virginia state is one of the first states to introduce laws banning and criminalizing the unlawful dissemination of falsely created material like digitally generated pornography known as deepfakes. (Virginia Legislative Information System, 2019)

- Texas state, in September 2019, criminalized the misuse of deepfakes through the Texas Senate Bill 751 (SB751) amendment to the state's election code. This law is enacted only in a political context. The act prohibits individuals to create and dissemination of deepfakes with an intent to harm a political candidate or manipulate the outcome of an election (Artz, 2019).
- Maryland has proposed a bill to legalize prohibition of the misuse of deepfakes similar to the laws of California in political context. (Lovells, 2020)

According to an article published by Shannon Reid, a graduate from University of Pennsylvania Law school, since its creation, deepfakes have always challenged the efficacy of U.S. Law in holding individuals accountable for their actions when they publish deepfakes of others without their consent. Unfortunately, there are no sufficient solutions for the targeted individuals in deepfake media in the U.S. federal and state laws. Privacy laws are insufficient in addressing the specific technologies and behaviors that pose the most significant threats. Furthermore, federal criminal and intellectual property statutes, which could potentially apply to deepfakes, are often narrowly interpreted by the courts or vulnerable to defenses that severely limit the legal options available to potential victims (Reid, 2021).

Current strategies to detect the deepfakes misuse:

At present, being a novel technology, the presence of standalone detection mechanisms exclusively designed to combat the misuse of deepfake technology are negligible. A comprehensive examination of the existing literature reveals a significant gap in the development of specialized tools and techniques solely dedicated to identifying and countering the potential malevolent applications of deepfakes. While research in the field of deepfake detection has made significant strides in detecting manipulated content, the focus has primarily revolved around identifying deepfakes in general without a specific emphasis on addressing their potential misuse. As the landscape of digital deception continues to evolve, it becomes increasingly imperative for researchers and technologists to direct their efforts towards the creation of dedicated measures that can effectively target the misuse of deepfakes, thereby bolstering the security and trustworthiness of media content in our digitally interconnected world.

Currently available methods to mitigate the Deepfakes misuse:

The methods to mitigate deepfake misuse have been continually evolving as the deepfake landscape advances, and a multi-faceted approach is often necessary to effectively address the challenges posed by deepfake technology.

One of the most popular social networking sites, Meta (formerly Facebook) acknowledges the severe consequences of non-consensual sharing of intimate images (NCII), often referred to as “revenge porn.” The company emphasizes its commitment to not allowing such content on its platforms and announces its efforts to combat the spread of NCII. Meta and Facebook Ireland, in collaboration with the UK Revenge Porn Helpline and more than fifty global organizations, have launched StopNCII.org. This unique platform is designed to offer a secure and confidential resource for individuals who are worried about the non-consensual sharing of their intimate images, which may include nudity or sexual content created using advanced AI technologies like deepfake to overcome and avoid sextortion (Antigone, 2021). The StopNCII.org website is designed to enable the individuals worldwide to take proactive measures to prevent and stop the unauthorized sharing of their private images on platforms like social media and other tech websites that are included in this initiative and there by provides greater control to the victims and enhances the security of their personal media.

There are numerous studies, reports, and research papers on detecting deepfake technology in the current situation of the cyberworld. A report generated on finding the effectiveness of deepfake on human attitude and intentions by conducting seven preregistered experiments in which groups of people have been exposed to the deepfake videos and genuine videos and tabulated the results on the levels of effectiveness of deepfake content on a viewer’s intentions and attitude. In their first experiment, they exposed the viewers to the genuine content which is a video of a novel individual in which he introduced himself and said few words about highly positive aspects and in another video, he spoke highly negative statements. A group of viewers watched the negative or positive videos and then completed corresponding implicit association test and the results from those tests showed that the genuine online content has strongly influenced the attitudes of the viewers and their intentions towards the person who spoke in the videos. This showed that genuine content can promote social learning at implicit and explicit levels.

In the second experiment, another group of people are exposed to the deepfake content, and it has been found that the change in attitude and intentions are same as the genuine content which is a concern to notice. In the fourth and sixth experiments they tried a different test data in which the group of people exposed to a deepfake content which was created from scratch and not from the prerecorded media which has been a successful attempt to control the viewers attitude and intentions.

In the third and fifth experiment the group of people acting as viewers are informed that they will be either exposed to genuine or deepfake content and yet their intentions and attitudes were changed even after being aware of the presence of deepfake. It means that deepfake has done its job even when it was an informed act and due to this it makes a deepfake technology highly dangerous as the damage it caused cannot be undone easily.

The paper published by (Westerlund, 2019) , highlights the possible threats of the deepfake technology which include major threats to the political and business system as well as the society involved with them. It makes the job stressful to the journalists to identify the real news from the fake ones and they may even give rise to the hard time trusting them. The study stated that there is a need for legislation and regulation to encounter the deepfake misuse and it suggests that the deepfakes are remarkable threat to the society, political and business systems and needs to be combatted by developing the deepfake detection systems and develop techniques to prevent the deepfakes and make the society ready to encounter the deepfakes. (Maras & Alexandrou, 2019)

According to an article published by, stated that it is not all the time the political deepfakes mislead the individuals, but it will implant a thought of uncertainty which will impact the trust levels on the news media. It may be still at initial state, however over a period, it will affect the online civic culture and is potential enough to produce provoking patterns among the viewers. Damage is not only measured in monetary terms but also psychological terms when it comes to deepfake misuse. They have conducted experiments to highlight the significance of the role played by the deepfakes in eroding trust in social discourse while contributing to the misinformation that is found online. They have taken large group of people from United Kingdom's population to gather the data required to analyze the people's assessment of deepfakes and found that the deepfakes makes people feel much more uncertainty than being misled and this uncertainty will in turn induce indeterminably and cynical behavior among the people and may even create trust issues on social media content and further create challenges in retaining the online civic ethics in democratic communities (Vaccari & Chadwick, 2020).

According to Nicholas O'Donnell, in one of his published articles, deepfakes, almost by definition are false and misleading in nature and the court has marked them to be unprotected. He has classified the deepfakes into two basic categories such as

- Deepfake pornography is the first category, which has an adverse effect irrespective of the number of viewers to the deepfake videos.
- The second category includes those deepfake media which requires certain level disclosure and distribution to have a destructive impact.

The threats which are caused by the deepfake content have mainly divided into three categories such as Elections manipulation, economic interference, and public safety issues. According to Nicholas, the legal procedures which are available currently are inadequate for confronting the threats associating with deepfake due to the presence of Section 230 of Communication Decency Act that supports the social media corporations in the cases where the accountability for the deepfakes that are spread virally on their platforms and the recent amendments do not provide much effect as the social media is immune from being liable to the information that is being posted and shared on their platforms. Even supposing that there is an amendment in place for the section 230, any rules applied to the deepfakes must be in compliance with the first amendment which would be a weak regulation as the deepfakes come under a category of video editing technology and also considered as a right to expression.

The above note is pertaining to the second category of the deepfakes and it also argues that the Section 230 of Communication Decency Act should be amended in such a way that it should consider the social media and online platforms as the disseminators of the deepfake content which is displayed on their applications which makes the federal regulations capable of penalizing the social media companies for the violation and can demand them to pay the fines for disseminating the false information like deepfakes. In most of the cases the fines issuance in large figures would be much more efficient than civil and criminal system of liability as this will force the organizations to remove or get rid of the offending deepfake media as soon as they are posted, and this kind of regulations can withstand the constitutional challenge.

However, the introduction of the Communications Decency Act (CDA) has provided the social media platforms with immunity against penalties and raised the issues at policy level and created constitutional impediments that any deepfake content will face and because of this, there is a need for the amendment to the Section 230 so that the social media platforms will be held liable (O'Donnell, 2021). Assistant Professor and a Deepfake pioneer, Hao Li says "*This is developing more rapidly than I thought. Soon, it's going to get to the point where there is no way that we can actually detect [deepfakes] anymore, so we have to look at other types of solutions.*" Talking about the rapid development of deepfake and the need to develop the combatting methods for the misuse of the deepfakes.

According to a study by Mika Westerlund (Westerlund, 2019), in which a review and analysis on 84 recent public news on deepfakes was done in order to assimilate the concept of deepfakes, who creates the deepfakes and its benefits and threats to the humankind and the current examples of the deepfakes and mitigative measures to contend them. The study has enabled them to know that the deepfakes are videos that are digitally maneuvered to look hyper-realistic so that they portray people in a way that they say and do the things that never occurred.

The study also found that, deepfakes are created using the GANs which can generate new content based on the existing data and these deepfakes of real people often tends to go viral and disseminates quickly on social media platforms and there by acts as an effective tool for the disinformation. The study also discovered the various contributions to the knowledgeable literature on the deepfakes which argues that deepfakes are usually promoted because of the dependency of citizens on the commercial media platforms, any deepfakes produced on heated conversations in the political context, usually false in nature can be easily disseminated online and the deepfakes gets its significance from the ability to utilize the advanced technologies like Artificial intelligence to produce hyper-realistic videos.

The study supports these arguments by indicating that commercial platforms, comprising both news media and social media platforms, are hotbeds for the production of deepfakes. These deepfakes are not solely based on heated political arguments but also on the broader social media context, leveraging the vast amount of online data available. This situation can erode trust in data shared on social networks and even lead people to doubt their prior beliefs.

The rise in the number of occurrences of fake news business models, which generate a significant amount of web traffic through advertisements, is well-documented in the study. This is supported by research analyzing news articles from journalists who sometimes rely on unethical techniques like clickbait.

Furthermore, the study identifies several factors associated with deepfakes. These include overly sensitive areas like governments, political extremists, lawbreakers, and vindictive individuals who create fake media to provoke online paid and unpaid trolls. Automated bots play a role in spreading this information on social media platforms. These actors are primarily motivated by intentions to harm others in several ways and to influence people to change their opinions on specific topics, leading to confusion in the public. Their ultimate goals may include financial gain or altering opinions about organizations, often for amusement or plain fun, which can impact individuals' lives.

Despite its valuable findings, the study has limitations. It analyzed a limited number of articles, specifically eighty-four online articles, to explore the concept of deepfakes. This number is relatively small given the ever-growing development of technology. More comprehensive insights might have been gained if a few more articles had been included in the analysis.

Secondly, the sources taken into consideration in the study included only the public sources like the online news sites for the article review and if other types data would have been included in the study like the online community discussions on deepfakes, considering the data gathered from the interviews given by the GAN developers and the artists in deepfake creation field where few of the artists are recognized as not only developers of the deepfake technology but also anti-deepfake technology developers who can provide further insights on the policies to be followed to overcome the issues of the deepfake and to mitigate the impacts of the deepfake in the lives of people. It also lacked the inclusion of opinions and views as the study did not include the commentary sections from the public news articles on the deepfake technologies so that the study could have analyzed the ideas of the readers. The study also could have formed the insights on how the deepfakes are perceived among the large audience so that there can be proper information available to work on the methods to combat the impacts of deepfake and can emphasize on the actual problem areas instead of beating around the bush. These impediments have paved the way for the new research in the field of deepfake as there is much more to be analyzed to overcome the issues of the deepfake.

Another study, Social Impact of Deepfakes by Dr. Hancock and Dr. Beilenson discusses the way researches are conducted in the field of deepfakes as it is a recent invention and still have scope in research despite the popularity of the deepfake media and technology like the Faceapp application and Zao App on social media and online platforms (Doffman, 2019). However, at the time of their research, only few studies have been taken into consideration which analyzed the aspects of psychological, societal and consequences of the policies that were in place at that time in a world where the media can be easily shared across the world instantly whose legitimacy is unknown in most of the cases and which are imperceptible to be real or fake (Hancock & Bailenson, 2021).

The need of their study was based on the face that there has been aplenty of research done on the methods to detect the deepfake but there aren’t much research and studies held on the impact of deepfakes on the society’s psychology and hence they come up with an idea to study and examine the social impact of the deepfakes and the possible impacts of deepfakes on the people. The study was mainly inspired by the Seitz’s and his colleague’s presentation on the sensational deepfake video of then President Obama, where a high-quality lip synching was achieved using deepfake where the mouth movements from the younger Obama was used to create a deepfake video of Obama twice his age which was perfectly drafted deepfake video with the help of machine learning techniques. (Suwajanakorn et al., 2017). There were two main things that were identified in the Seirz presentation which were discussed in this study which are as follows:

- The aspect was the fact that the algorithms used to generate deepfakes are much easier than the algorithms used to detect them as per the basic nature of the GANs. The deepfakes have been noticeably migrated from the computer programs in science Laboratories to mere mobile apps which made it easy to generate a deepfake content by a commoner who do not need much expertise in the field of deepfake creation. Thus, the creation of deepfake has much advantage over the detection of deepfakes.
- The second aspect was the social and psychological impacts of deepfake and are there any proper studies made on this aspect of deepfakes.

These were the main aspects that led to the initiation of this study. The study discussed the concerns in the field of research on the deepfake technology such as:

1. Scarcity of empirical research: At the time of this study being conducted, there were ample studies made on the creation of methods to detect the deepfakes. However, there were not many studies and research made on the social impact of the deepfakes where there exists rapid dissemination of deepfakes in the digital world which are usually indistinguishable from the real videos. Although there were various studies held on social prejudices and memory attainment from the altered still photos, the psychological consequences of watching the AI crafted videos have remained substantially remain unstudied. Peculiarly, the virtual reality has been a great starting point for studying the social impact of the deepfakes. In virtual reality, there is a feature where doppelgangers in the form of 3D models of a given individua from the 2D images and once the doppelgangers are built, its much simpler to generate the animations of this 3D models and rendered as the 2d videos which looks indistinguishable with the real videos. Watching VR videos can cause false memories which makes the individuals believe that they have performed the deepfake activity and can even influence their preferences on the brands the doppelganger uses in the VR (Segovia & Bailenson, 2009).
2. Few observations from the deception research: The Fundamental nature of the deepfakes is deception which involves intentional, deliberate misleading of an individual. The Study suggests that according to the literature on deception, people tend to believe false evidence and are not usually good at detecting deception while reading the messages. Also, studies have shown that the level of

deception detection is almost the same as in the case of messages and audio messages or video clips.

The study found that it is usually surprising that accuracy of detecting deception is much low as the people tend to believe what others convey without much effort. Also, it is found that deception happens much easier when visual media like videos as humans rely much on visual aids while perceiving information. Although people easily believe the deception in the form of videos, knowing the fact that deepfake videos exist might change their perception further as it may create a doubt on believing the video contents to be real and it may certainly interfere with the ability to acquire knowledge in the world of deepfakes. This may in turn undermine the role of journalism and other media in the current digital world.

3. Aftermaths of the Deepfakes: According to an empirical study by Vaccari and Chadwick (Vaccari & Chadwick, 2020), it is found that there is a rise in the sense of uncertainty when the individuals have to trust the news as they are aware of the existence of deepfakes similar as in the case of spam emails where the individual easily ignores the email after becoming aware of such scams.

   Hence, viewers are found to be developing resilience to the new forms of deception like deepfakes. One of the most important concerns of using deepfakes is the nonconsensual victim being portrayed in the videos or images usually as in the case of pornography alterations in which they have never engaged. The impact of such incidents can be devastating on the lives of the victims as the main motivation behind their creation is they will be mainly used to humiliate, extort, or harass the victims.

4. Future scope for the research: This study urges the researchers to conduct studies on the social impacts caused by deepfakes as there is a remarkable development in this field and can produce some interesting insights from the research as the current study although being exploratory, but it is just preliminary. The study also indicates that there is a need for the attention on frontier which is usage of the AI powered filters that allows modification of the videos in real time usually the act of smiling which shown positive effects on the viewers and stopping them to detect the deepfake effects most of the time. The deepfakes, despite being able to undermine the trust in the media and falsely manipulate the attitudes of the society, it also turned into a more common place to use deepfakes in communication context in daily activities and the study makes it clear that there is need for conducting much more empirical research on the social and psychological impacts of the deepfake as it is evolving like never before. (Hancock & Bailenson, 2021)

Summary

This chapter meticulously presents the context of the topic through background delineation. It accentuates issues related to the mishandling and abuse of deepfake technology, highlighting their severity. The chapter includes a detailed literature review on existing legislation on deepfakes across the globe and underscores the need for awareness on these issues. Furthermore, it discusses the extent of detrimental incidents that have occurred in the past. Additionally, the chapter describes literature related to existing methodologies in the current topic.

Chapter III: Methodology

Introduction

The design of the study for the current topic is elaborated in a detailed manner in this chapter including the methods followed for data collection. I have executed the designated set of activities outlined in the Design Study to map out a structured method, carefully crafted to concentrate, gather, and meticulously investigate supplementary material on the subject.

Design of the Study

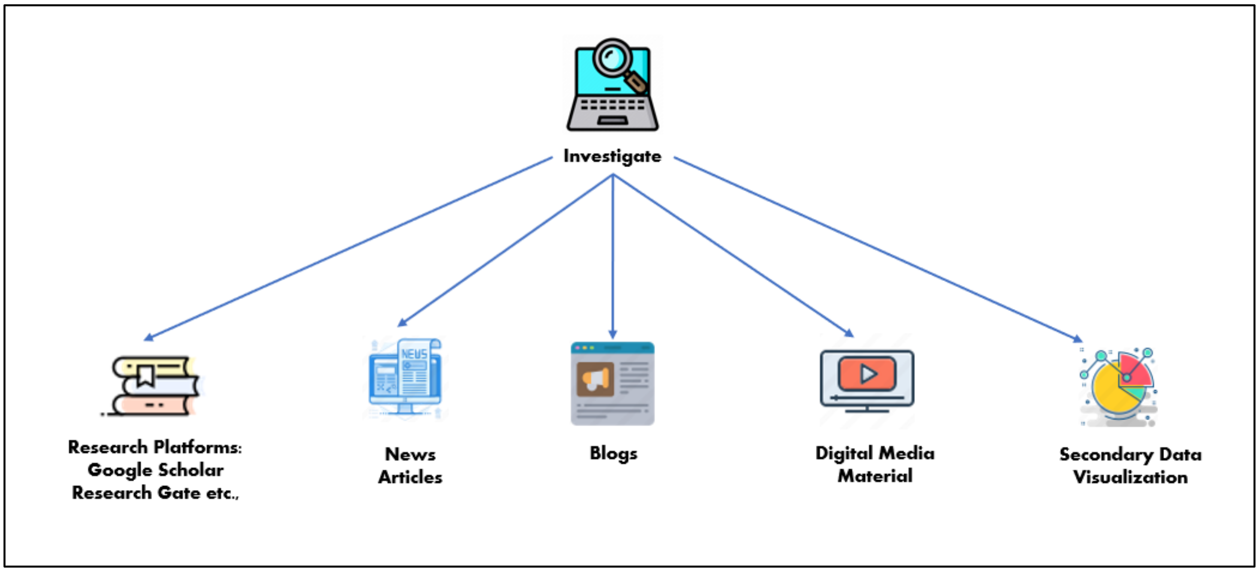
1. Research on the emergence of deepfake technology using online research platforms like Google Scholar, SCSU Library and Research Gate.
2. Investigate the recent works to get the latest advancements and trends in deepfake technology usage and its abuse.
3. Analyze the researched data and document in an organized way to present it to the readers in a comprehensive way.
4. Delineate the pressing need for increasing awareness in a thematic approach and suggest a technique to enhance awareness and hinder the losses that can be caused by the misuse of deepfake technology among the society.

Data Collection

The process of data collection comprises of a comprehensive literature review conducted on online research platforms, including Google Scholar, SCSU Library, ResearchGate, and various miscellaneous articles from reputable websites and blogs. The data collection process is thorough and wide-ranging. It draws heavily on scholarly journals in the realms of technology, law, and ethics, yielding foundational understanding and tracing the legal development of deepfake technology. The mission is to analyze how policymakers confront the difficulties posed by deepfakes and to see how these responses change as technology strengthens. Furthermore, it delves deep into industry reports, tapping into the knowledge of IT professionals and specialists. These sources provide insight into the practical elements of deepfakes such as their formation, discovery, and the approaches that persons and companies are utilizing to safeguard themselves. These reviews, often, carry data-driven penetrations, market evaluations, and forecast projections that are crucial for grasping the present situation and foreseeable future of deepfake technology. This research also draws from an immense stockpile of digital media materials, such as research papers, tech blogs, and forums where experts and intellectuals converse regarding the forefront of deepfake technology.

Figure 1

Data Collection


Investigate
Research Platforms: Google Scholar Research Gate etc.,
News Articles
Blogs
Digital Media Material
Secondary Data Visualization

Summary

This chapter provides a comprehensive summary of the research methodology and study design. It outlines the data collection procedures. The chapter also gives an overview of the data collection process to measure the level of awareness on deepfake technologies. Furthermore, it delineates data analysis procedures, focusing on identifying and presenting the collected data.

Chapter IV: Analysis and Discussion

Introduction

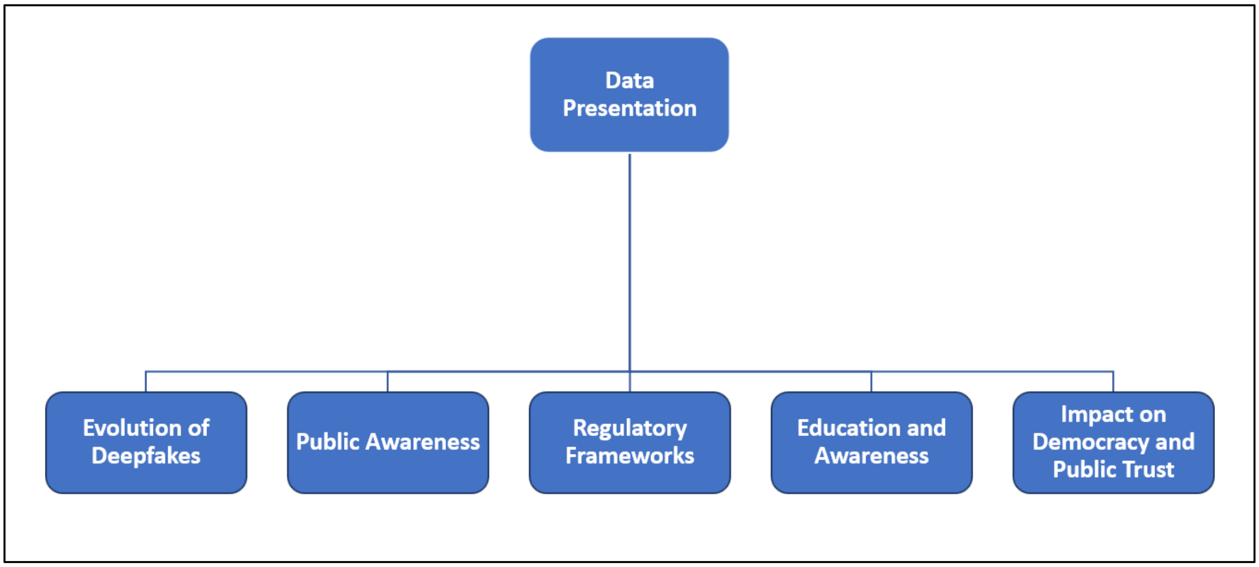
This chapter seeks to prove if the research questions mentioned in the methodology section have been completely addressed. It pays special attention to the advancement of deepfake technology, its improper uses, the various legal retorts it has bred, and the level of comprehension citizens have of its results on privacy and security. The findings have been systematically organized and the results are organized into thematic categories to make the data easier to comprehend. This effective categorization enables a clearer understanding of the multiple facets of deepfake technology. Furthermore, this chapter will examine emerging trends, interpreting their correlation to the discourse surrounding deepfake technology and its far-reaching consequences. In its essence, this chapter functions as both a storehouse for findings and an infrastructure for comprehending the delicate link between the progression of deepfake formation technology and ensuing issues of secrecy and safety. It follows a carefully planned format to help readers advance through the tangled circumstances in play, equipping them with a multi-level comprehension of the trials and effects that deepfakes bring to the electronics age.

Delineating the Deepfakes and its Implications on Society and Individuals

The research conducted concerning deepfake technology and its consequences for security and privacy is presented in a thematic manner, with meticulous care, pays attention to transparency and perceptibility. The objective is to offer a lucid and comprehensive account that not only conveys intricate knowledge but also fascinates the reader’s concentration, consequently making the implication of this technology unambiguous. To this objective, visual aids assume a paramount role. Charts segment the emergence of deepfake technology within different fields.

Figure 2

Data Presentation



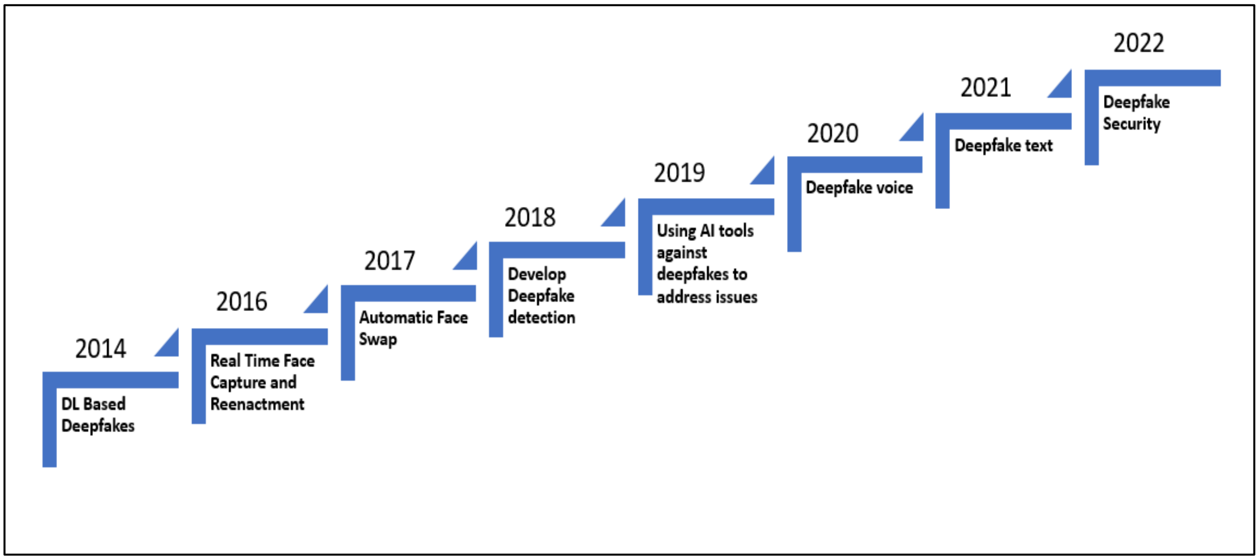
As shown in figure the main themes of this data include Evolution of deepfake technology, public awareness and implications, regulatory frameworks, education and awareness, impact of deepfakes on democracy and public trust.

Evolution of Deepfakes:

Deepfakes have advanced significantly since their inception, owing much of their development to the field of artificial intelligence (AI). Deepfakes evolved with rapid technological advancements in less than a decade as shown in figure 3.

Figure 3

Historical Development of Deepfake Technology


2014
DL Based Deepfakes
2016
Real Time Face Capture and Reenactment
2017
Automatic Face Swap
2018
Develop Deepfake detection
2019
Using AI tools against deepfakes to address issues
2020
Deepfake voice
2021
Deepfake text
2022
Deepfake Security

A brief overview of the rapid evolution of deepfake technology through years from 2014 to 2022 is as follows:

2014: The term deepfake was first introduced by researchers using GAN AI and provided technical basis for the deepfakes, one of the major advancements in AI. During this time, researchers combined GANs with CNNs for efficient image recognition using parallel processing aimed towards credible results. (Nguyen et al., 2022)

2016: Development of face capture and reenactment using combination of GANs and parallel processing to modify the AI learned data showing credible results in creating fake portraits. The first occurrence of viral deepfake video. (Thies et al., 2016)

2017: Major development in the quality of Deepfakes by rectifying the major drawbacks in the previous GANs. Stage wise training of GAN network has enabled the creating of high quality deepfakes with minimal flaws which looked more convincing to believe as real images of people.

2018: Improved GAN control has enabled developing more realistic images of the target. Rapid development of deepfake media in the form of videos and images including majorly pornographic content which has obliged the researchers to develop Deepfake detection techniques.

2019: Deepfake has found its presence in the mainstream. Researchers from Samsung have developed GAN to create deepfakes on humans and artwork. Israeli researchers introduced face swapping AI which can swap the faces in real time with no prior training by just using the application (Radford et al., 2016). The policy makers step in to hinder the misuse and rapid dissemination of deepfake content to protect the privacy and security of the victims. Few governments across the globe including USA, China and Germany have taken measures to address the Deepfake related risks in a society and political arena.

2020: Evolution of Deepfake content is at rise by creating deepfake audio using Voice clones which can be both a challenge and an opportunity in the field of technology. The AI organizations started actively developing advanced detection software. These detection tools utilize cutting-edge AI algorithms and machine learning models to analyze media content and identify signs of manipulation or falsification (Schreiner, 2022).

2021: Development of Deepfake text using GPTs which allows generation of fake content and like news and articles which can actively aid in frauds and scams. This technological advancement allowed the creation of extremely persuasive fake news articles and written content that can closely emulate the writing style and tone of genuine individuals.

2022: Development of Detection techniques for mitigating the Deepfake misuse using CNNs and online platforms to report and remove the unauthorized deepfake content (Ahmed et al., 2022).

Public Awareness:

Public consciousness regarding deepfakes is often inconsistent, usually due to contemporary events or press coverage. Those with limited computer knowledge are particularly vulnerable, thus emphasizing the requirement for comprehensive informational campaigns specifically tailored to reach and teach this demographic.

The awareness levels of public on IT and Internet, according to an article on the ability of respondents to correctly detect deepfake and authentic videos by Doss and his group (Doss et al., 2023), the diverse group of people including adults, principals, teachers and students are vulnerable to deepfake videos as many of them are unable to identify the authenticity of the content. It indicates that this vulnerability has critical implications for the current education system, as it could lead to the dissemination of misinformation among students, potentially affecting their views on science and policy.

The study suggested that while technical aspects like video quality play a role in detecting deepfakes, analyzing social aspects, such as speaker knowledge and content credibility, is equally important. Moreover, older individuals and those with high trust in information sources are more vulnerable, indicating a need for tailored educational interventions and traditional media literacy approaches.

The trajectory of deepfake technology at present paints a stark picture of the contrast between quickly advancing technology and the lag in public awareness. This mismatch brings significant threats to privacy and safety because of the ability of deepfakes to be deployed for malicious purposes. Even though phony versions of deepfakes are showing up more often and being utilized in misinformation campaigns, the overall public's cognizance of the ephemeral nature and implications of these detailed counterfeits remains lacking. Studies and research have shed light on the pervasive disparity in knowledge, exhibiting that even as deepfakes become more convincing and harder to identify, the public's capability to differentiate them from the original media does not scale accordingly. This discrepancy renders individuals and organizations susceptible to manipulation, with potentially far-reaching effects ranging from personal reputation hurt to geopolitical upheavals.

This study by Doss shows significance of addressing the social context of deepfake media and highlights the challenges posed by advancing deepfake generation technologies. In the long term, the study underscores the need for education on deepfake detection to combat the spread of misinformation effectively. To guarantee such initiatives are accessible and efficacious for those not accustomed to technological advances, sound communication tactics must be formulated.

Regulatory frameworks:

The complexity of identifying deepfakes is heightened by the reality that once new detection techniques are developed, corresponding progressions in deepfake creation can collapse these recently formed obstructions. This back-and-forth struggle implies the need for steady awareness and adjustability in identification technologies. Furthermore, the efficacy of deepfake detection tools appears to be highly dependent on the type of media or platforms they are applied to and their capability to be used on a large scale. An active stance must be taken, necessitating a collective effort to bring together the versatility of startups and the comprehensive resources of established entities. This implies that government bodies, academic establishments, and industry trailblazers must join hands to aid ongoing investigations.

The intrusion of deepfake technology carries substantial risks across numerous areas, such as safety, politics, and civil liberty. The likelihood of deepfakes being abused to devise fictitious stories, mimic individuals, and circulate untruths is exceptionally high. Therefore, it is essential for authoritative agencies to intervene and develop a resilient legal system capable of shielding individuals and civilizations from these detriments. Without suitable controls, deepfakes have the potential to erode public faith in electronic correspondence, which underpins this day and age's communication. Such a change could have a wide-reaching result, affecting areas from reporting to court proof, in which the reliability of sound and visual accounts is of the utmost importance.

Furthermore, the psychological effects that deepfakes have on victims, which can vary from a blow to their reputation to emotional anguish, should not be disregarded. In the sphere of politics, deepfakes can be utilized as an instrument of skullduggery to control electoral outcomes, shake confidence in public figures, and disrupt democratic systems. The urgency of global collaboration in tackling this matter is self-evident. It is generally agreed upon that deepfakes show no respect for political boundaries; as such, a collective strategy of action may be the best approach to regulating them.

For this to be accomplished meaningfully, the alignment of laws between nations is fundamental in order to avoid malicious entities finding havens of refuge. In order to achieve desired ends, it is essential for governing bodies, industry players, and citizens to collaborate with each other to set proper parameters for the fabrication and dissemination of deepfake content. The advancement of deepfake tech has made remarkable steps forward in terms of generation and identification possibilities. Even though there are numerous potential applications, most notably in the entertainment industry, this technology also carries considerable risk, most notably its power to propagate false information and influence collective opinion.

The absence of stringent regulations in the realm of deepfake technology highlights a pressing need for comprehensive oversight. With limited legal frameworks in place, the potential for misuse and abuse of deepfake capabilities remains largely unchecked. This regulatory gap has profound implications for various sectors, including media, cybersecurity, and privacy. The lack of specific guidelines and standards exposes individuals and organizations to substantial risks, from reputational damage and privacy breaches to national security threats.

The imperative for a comprehensive regulatory framework addressing deepfake technology is paramount in today's digital landscape. Deepfakes, driven by advanced AI algorithms, have the potential to disrupt the very foundations of trust and authenticity in various domains, including media, politics, and business. As they become increasingly sophisticated, deepfakes can deceive individuals and manipulate public opinion, posing serious threats to national security, privacy, and the credibility of institutions. The importance of such a regulatory framework lies in its capacity to set clear guidelines, standards, and legal boundaries for the creation, dissemination, and detection of deepfake content. It provides the necessary tools to combat the malicious use of this technology, ensuring transparency, accountability, and ethical considerations in its applications. Moreover, a regulatory framework can foster international collaboration, harmonizing efforts to mitigate the global impact of deepfakes. In essence, the establishment of a robust regulatory framework is essential to navigate the challenges posed by deepfakes, safeguard the integrity of information, and uphold the principles of truth and authenticity in an increasingly digital and interconnected world.

Regulations play an important role in addressing the risks associated with deepfakes. Initiatives such as the development of new technologies to detect deepfakes, the promotion of media literacy and critical thinking skills, and the establishment of regulations and standards to govern the use of deepfakes in various contexts to ensure that people understand the risks associated with deepfakes and can identify them when they encounter them.

Education and Awareness:

The importance of education and awareness of deepfake technology cannot be overstated in the contemporary digital age. Deepfake technology, enabled by powerful artificial intelligence algorithms, has ushered in an era where the lines between reality and fabrication are increasingly blurred. These sophisticated algorithms can generate hyper-realistic audio, video, and written content, making it nearly impossible for the average individual to discern fact from fiction. The implications of this technology are far-reaching and have consequences across various sectors, including politics, journalism, entertainment, and, importantly, the broader societal landscape.

Teaching media literacy and critical thinking in schools and universities and High-risk professions like journalism and politics could greatly benefit from training on media verification. Law enforcement and legal professionals might require training in collecting digital evidence and identifying deepfake use in criminal activities.

First and foremost, education and awareness are crucial in safeguarding individuals and society at large from the pernicious effects of deepfakes. By imparting knowledge about the existence and capabilities of deepfake technology, individuals can become more discerning consumers of digital content. They can develop the critical thinking skills necessary to question the authenticity of the information they encounter. Moreover, education empowers individuals to identify potential red flags, such as inconsistencies in content, unnatural facial movements, or suspicious sources, which are telltale signs of deepfakes.

Awareness campaigns are vital for reaching a wider audience, including those without formal education on deepfake technology. These initiatives raise public awareness about deepfakes and its potential uses, fostering vigilance. Besides protecting individuals from misinformation, education and awareness are essential for upholding democratic integrity, especially during elections, and preserving trust in institutions. Deepfake content can undermine public trust, making education a vital tool in combating this erosion. Additionally, education and awareness drive the development of mitigation strategies and technologies by understanding deepfake creators' methods. This collaborative approach is crucial to outpace malicious actors exploiting this technology. In essence, education and awareness about deepfake technology are paramount in the information age, protecting individuals, democracy, and fostering digital resilience. As deepfake technology evolves, our efforts to educate the public about its existence and potential consequences must evolve in tandem.

Impact of Deepfakes on Democracy and Public Trust among the Society

Another key aspect to consider is the potential impact of deepfakes on democracy and public trust. Deepfakes have the potential to undermine public trust in media and information sources, which could have significant implications for democratic societies (Chesney & Citron, 2019). As such, it is important to consider how deepfakes might be used to manipulate public opinion and how to develop effective strategies to combat such activities. To understand the multifaceted nature of this issue, we need to delve into the various dimensions of its impact.

Misinformation and Manipulation: Deepfakes can be used to create deceptive content that mimics the appearance and speech of real individuals. This raises concerns about the manipulation of political figures or candidates, where fabricated speeches or statements can mislead the public and influence elections. Such manipulations can undermine the core principles of democracy, which rely on informed and free choices by citizens.

Public Trust and Media Credibility: The widespread dissemination of deepfake content can lead to a decline in public trust in media and information sources. If people become skeptical of the authenticity of visual and audio content, they may question the credibility of news outlets and sources they once relied on. This erosion of trust can have far-reaching implications for the functioning of democratic societies, as informed citizens are essential for making sound decisions and holding institutions accountable.

Policy Implications: The presence of deepfakes poses policy challenges. Governments and regulatory bodies need to consider how to address this technological threat. Crafting effective legislation to curb the creation and dissemination of malicious deepfakes without infringing on freedom of speech is a delicate balance that policymakers must strike.

Furthermore, it is important to consider the broader context of technological change and innovation in relation to deepfakes. As deepfake technology continues to evolve, it is likely that new and more sophisticated forms of manipulation will emerge, making it increasingly difficult to detect and combat the use of deepfakes. As such, it is important to continue to invest in research and development to stay ahead of the curve and to ensure that societies are equipped to address emerging challenges associated with deepfakes and other emerging technologies.

Methods used for Analysis

The analysis of this research is carried out as an elaborate amalgamation, combining various origins to note similarities, observe tendencies, and recognize deficiencies among the considerable corpus of deepfake technology. This is a fundamental process, making sense of the collected data and furnishing an organized appraisal of the privacy and security aspects that are involved. Using a thematic approach, the information can be segregated into logical classifications, which are essential for comprehending the intricacies of deepfake technology. Every theme reflects a significant part of the overall narrative. The evolution of deepfakes chronicles the swift growth of deepfake generating and identification techniques. The literature related to the problem discusses the occurrences where deepfakes have been used against other persons and corporations. Regulatory Frameworks characterizes the international legal field as it fights to remain abreast of technological progress. Public awareness discusses collective awareness and concerns regarding deepfakes. This analysis is more than just an amalgamation of the research data. Rather, it forms a vital instrument to compute the necessity and magnitude of problems aroused by deepfakes.

By carefully working out details like the pace of technical development, the quantity of deepfake transmission, and the response times of governing bodies, the exploration forms a crystal-clear representation of the current conditions. The thematic architecture facilitates contextualizing the reviewed data, guaranteeing that the story is both exhaustive and widespread. The objective is to furnish a comprehensive yet intricate overview of the milieu, allowing onlookers to apprehend the specifics of every thematic domain while recognizing their interrelationship and the general conclusions for policy, law, and social conventions. In summary, the data analysis is a delicate endeavor, considering the immediate risks and issues composed by deepfake technology while simultaneously positioning them within an encompassing thematic structure that speaks to the advancing dynamics of digital confidentiality and safety. This comprehensive analysis is essential for stakeholders who must negotiate and devise answers to the multifarious challenges presented by the proliferation of deepfakes.

Chapter V: Conclusion and Recommendations

Introduction

In this chapter, I will summarize the methodology employed to study the Privacy and Security Implications of Deepfake Technology dissemination, conclusion of the study, recommendations, and future work. Additionally, I will verify whether the research questions outlined during the methodology phase have been adequately addressed in the study.

Results

The study highlighted the potential impact of deepfakes and the scenarios where they could have the most significant consequences. Range of potential scenarios, including political campaigns, celebrity scandals, identity theft and financial fraud where deepfakes can be employed to spread misleading or false information, raising concerns about the credibility of information, particularly in political contexts and media; possibility of invasive privacy breaches impacting the personal security and well-being; have been discussed in a comprehensive manner.

Overall, addressing deepfake risks requires a comprehensive approach, encompassing prevention, regulation, education, and public awareness. This multifaceted strategy can essentially mitigate the potential harms of this emerging technology. Additionally, ethical, and legal considerations related to deepfake creation and dissemination, including issues of privacy, consent, and intellectual property, should be addressed.

After conducting the research throughout this study, the main objectives of this study that are defined during the methodology have been achieved and addressed below.

1. Research on the emergence of Deepfake technology using online research platforms like Google Scholar, SCSU Library and Research Gate.

    The above-mentioned objectives of the study have been achieved by conducting extensive research on the emergence of Deepfake Technology, its usage and potential implications on society in terms of privacy and security issues that are caused by the dissemination of deepfakes, by utilizing online research platforms like Google Scholar, SCSU Library, and Research Gate various steps has been taken.

2. Investigate the recent works to get the latest advancements and trends in deepfake technology usage and its abuse.

    Various relevant articles and publications were identified by searching on these platforms with the relevant keywords like "deepfake technology", "fake videos" and "digital manipulation" etc. Furthermore, to refine the data for the study, a date range and publication type filter is applied during the search for the relevant information.

3. Analyze the data and document in an organized way to present it to the readers in a comprehensive way.

    A thorough study has been done on the abstracts and full texts of the selected articles on the deepfake emergence, its usage and potential implications are reviewed to determine their relevance and potential to contribute to the study. The sources that were deemed relevant were included in the literature review section of the study. In addition to the research platforms like Google Scholar, SCSU Library and Research Gate that were used to gather information for this study, other sources were used to conduct research on the emergence of deepfake technology including news articles, government reports and other media resources that provided the information on the major developments and use of deepfake technology, related legislature and current mitigative measures to combat misuse of deepfakes. The combination of various sources allowed for a comprehensive analysis of the emergence of deepfake technology and its potential implications for society and measures to be taken in case of misuse of the deepfakes.

4. Delineate the pressing need for increasing awareness and suggest a technique to enhance awareness and hinder the losses that can be caused by the misuse of deepfake technology among the society.

The study culminated in the development of a comprehensive plan to lessen the damaging impacts of erroneous deepfake exploitations, incorporating initiatives for digital savvy, the promotion of automated exposure technologies, revised laws to take action against misuses, and amalgamation between different sectors by categorizing the training groups. This multi-pronged approach, based on the rigorous inspection of interdisciplinary sources, counsel from experts, and thematical examination, demonstrates the effectual fulfillment of the study's objectives via a well-constructed blueprint, which endeavors to strengthen societal protection from the sly dangers posed by deepfake technology.

Conclusion

This paper presents and overview of deepfakes and discusses its societal impact on security and privacy, highlighting its challenge in discerning real media from fake media. Furthermore, the emergence of deepfakes poses significant societal and business challenges, potentially harming individuals, and eroding media trust. This paper has also highlighted the necessity for regulations, detection, mitigation, and prevention measures against deepfake misuse. The opportunities available to cybersecurity and AI to combat fake news and media have been discussed and it emphasized the need for addressing and increasing awareness on potential dangers of deepfakes among the policymakers and public. The obtained data on Deepfake awareness among have been analyzed to classify the audiences into different groups of varying awareness levels on deepfake technology. Suggestions have been made to address the lack of awareness among the target audiences by providing possible and the need for the effective ways to overcome the lack of awareness of this novel technology among the society and a future scope has been defined to lay a path for the future possible research on the implications of deepfakes on the security and privacy of the individuals in a society.

Although this project may not solve all the issues of deepfake technology, it aims to spread awareness and equip the public with knowledge to defend themselves against it. The long-term solution involves critical thinking and research, aided by technology. The project has aimed to contribute to this solution by exploring the power of design to reveal the truth and create awareness.

Recommendations

1. To address the deficiency in knowledge, concerted steps should be taken to ensure the development of education and public engagement. These initiatives ought to have a focus not only on the spotting of deepfakes but also on fully comprehending their wider repercussions on society. If citizens are not adequately educated and are not able to analyze media in a critical manner, then the opportunity is still present for deepfakes to be exploited for malevolent aims, which includes the breakdown of confidence in digital correspondence, an essential pillar of contemporary communication. In response, this research proposes the organization of comprehensive educational programs and efficient communication methods, aiming to advance digital literacy among numerous members of society. As technology advances, individuals must learn to effectively contend with the problems it poses, stressing the requirement for consistent, adaptive training initiatives to ensure individual and collective safety within the digital landscape

   As the target audience includes people from diverse backgrounds, the type of training that can be given will depend on the kind of target audience and their level of knowledge and skills. Upon thorough examination of potential target audience categories, the training pertaining to deepfake awareness can be methodically classified into five primary categories as described below:

   a. General awareness training: General awareness training should be provided to individuals with limited knowledge about deepfakes. This type of training would provide an overview of deepfakes, including their creation process and the possible consequences they could have on both individuals and society.

b. Detection training: This training can be beneficial for individuals who are responsible for verifying or authenticating media content, such as journalists or social media moderators. It could cover topics such as how to detect deepfakes using various techniques, tools, and technologies.

c. Prevention training: Individuals who are involved in creating or sharing media content, such as social media users or content creators, could benefit from prevention training. This training would provide guidance on how to prevent the creation or spread of deepfakes by using secure authentication methods or watermarking technologies.

d. Legal training: Legal professionals such as law enforcement officials and policymakers, as well as individuals in the legal field, could benefit from legal training on deepfakes. It could cover topics such as the legal implications of deepfakes, including issues related to privacy, intellectual property, and defamation.

e. Ethical training: Individuals who produce or consume media content, policymakers, and educators could benefit from ethical training on deepfakes. This training would focus on ethical considerations related to deepfakes, including topics such as consent, manipulation, and bias.

2. Recommendations to AI Industry: The field of Artificial Intelligence (AI) holds immense significance and impact across various industries and societal domains, and it is the fundamental aspect of the rapid development of deepfakes. The AI

specialists must invest in robust detection technologies for deepfakes as the misuse of deepfakes is inevitable and hence there is always a requirement for robust detection mechanisms to mitigate and prevent deepfake misuse. In addition to this, AI researchers shall collaborate with industry professionals and policymakers to address the deepfake challenges and advocate the establishment of standards and best practices. Furthermore, AI industry can work towards the establishment of continuous monitoring of Deepfake trends and adaptation of countermeasures as it is rapidly evolving technology.

3. Conduct a larger-scale survey**:** Currently, there are very smaller number of empirical studies on awareness levels on Deepfakes, so future work could involve promoting conducting extensive surveys to obtain a broader range of perspectives and a more representative sample to analyze the public awareness of deepfakes. This approach enables researchers to explore the evolution of awareness over time, capturing any shifts in public perception or knowledge regarding deepfakes.
4. Developing the legislation at federal government level: The need for legislation at federal government level on deepfakes can be addressed by developing the comprehensive and flexible legislation, along with international collaboration considering the various factors like severity of the perceived threat, inadequacy of the existing laws and as well as societal and ethical aspects.
5. Investigate the effectiveness of deepfake detection tools**:** As deepfake technology evolves, there is a need to develop effective detection tools to

mitigate the risks associated with their use. Future work could investigate the effectiveness of various deepfake detection tools and how they can be improved.

6. Explore the ethical and legal implications of deepfake technology: Deepfake technology raises a range of ethical and legal issues, such as privacy violations, identity theft, and defamation. Future work could explore these issues in more detail and identify ways to address them.
7. Identify and mitigate the social and political impacts of deepfakes: The dissemination of deepfakes can have serious social and political consequences, such as misinformation, manipulation, and interference in elections. Future work could focus on identifying these impacts and developing strategies to mitigate their effects.

Overall, this study can be used to inform the policymakers, educators, and the public about the potential harms of deepfake technology and the need for greater awareness and mitigation strategies for the security and privacy issues caused by the dissemination of the deepfakes in the current AI driven world.

References

Ahmed, S. R., Sonuç, E., Ahmed, M. R., & Duru, A. D. (2022). Analysis Survey on Deepfake detection and Recognition with Convolutional Neural Networks. *International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA).* IEEE. Retrieved from https://ieeexplore.ieee.org/abstract/document/9799858

Antigone, D. (2021, December 2). *Strengthening Our Efforts Against the Spread of Non-Consensual Intimate Images*. Retrieved from about.fb.com: https://about.fb.com/news/2021/12/strengthening-efforts-against-spread-of-non-consensual-intimate-images/

Artz, K. (2019, October 11). *Texas Outlaws "Deepfakes"—but the Legal System May Not Be Able to Stop Them* | *Texas Lawyer*. Retrieved from www.law.com: https://www.law.com/texaslawyer/2019/10/11/texas-outlaws-deepfakes-but-the-legal-s

Bregler, C., Covell, M., & Slaney, M. (1997). Video rewrite: Driving visual speech with audio. *Proceedings of the 24th annual conference on Computer graphics and interactive techniques.* (pp. 353-360). ACM Press/Addison-Wesley Publishing Co.;.

Briscoe, S. (2023, July 24). *U.S. Laws Address Deepfakes*. Retrieved from Asisonline.org: https://www.asisonline.org/security-managementmagazine/latest-news/today-in-security/2021/january/U-S-Laws-Address-Deepfakes/

Burchard, H. V. (2018, May 21). *Politico*. Retrieved from https://www.politico.eu/: https://www.politico.eu/article/spa-donald-trump-belgium-paris-climate-agreement-belgian-socialist-party-circulates-deep-fake-trump-video/

Chesney, B., & Citron, D. (2019). Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *CALIFORNIA LAW REVIEW*, 1779-1783.

Chesney, R., & Citron, D. K. (2018, July 14). Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *SSRN*, p. 1785.

Çolak, B. (2021, January 19). *Legal Issues of Deepfakes*. Retrieved from internetjustsociety.org: https://www.internetjustsociety.org/legal-issues-of-deepfakes#:~:text=However%2C%20deepfakes%20have%20some%20drawbacks,rights%20and%20intellectual%20property%20rights.

CVISIONLAB. (n.d.). *Deepfake (Generative adversarial network)*. Retrieved from www.cvisionlab.com: https://www.cvisionlab.com/cases/deepfake-gan/

Doffman, Z. (2019, January 27). *Chinese deepfake app ZAO goes viral privacy of millions ‘at risk’*. Retrieved from https://www.forbes.com/: https://www.forbes.com/sites/zakdoffman/2019/09/02/chinese-best-ever-deepfake-app-zao-sparks-huge-faceapp-like-privacy-storm/?sh=644d2b2e8470

Doss, C., Mondschein, J., Shu, D., Wolfson, T., Kopecky, D., Fitton-Kane, V. A., . . . Tucker, C. (2023). Deepfakes and scientifc knowledge dissemination. *Scientific Reports*.

Fagni, T., Falchi, F., Gambini, M., Martella, A., & Tesconi, M. (2021). TweepFake: About detecting deepfake tweets. *Plos.org*. Retrieved 2023, from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0251415

Hancock, J. T., & Bailenson, J. N. (2021, March). The Social Impact of Deepfakes. *Cyberpsychology, Behavior, and Social Networking*, pp. 149-152.

Hughes, S., Fried, O., Ferguson, M., & Hughes, C. (2021). Deepfaked online content is highly effective in manipulating people's attitudes and intentions. *TBD*, 2-4.

Jia, Y., Zhang, Y., Weiss, R. J., Wang, Q., Shen, J., Ren, F., . . . Wu, Y. (2018). Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (pp. 4480-4490). Curran Associates Inc.

Lovells, H. (2020). *Deepfakes: An EU and U.S. perspective.* Retrieved from GMCQ_-_Spring_2020_Deepfakes: https://f.datasrvr.com/fr1/320/16758/1207330_-_GMCQ_-_Spring_2020_Deepfakes.pdf

Maras, M.-H., & Alexandrou, A. (2019). Determining authenticity of video evidence in the age of artificial intelligence in the wake of Deepfake videos. *International Journal of Evidence & Proof*, 255-262.

Nguyen, T. T., Nguyen, Q. V., D. T., Nguyen, D. T., Huynh-The, T., Nahavandi, S., . . . Nguyen, C. M. (2022, October). *Deep learning for deepfakes creation and detection: A survey. Comput. Vis. Image Underst.* Retrieved from https://www.sciencedirect.com: https://www.sciencedirect.com/science/article/abs/pii/S1077314222001114#preview-section-abstract

O'Donnell, N. (2021). *Have We No Decency Section 230 And The Liability Of Social Media Companies For Deepfake Videos.* University Of Illinois Law Review.

Petkauskas, V. (2021). *Report: number of expert-crafted video deepfakes double every six months.* CyberNews.

Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised Representation Learning With Deep Convolutional Generative Adversarial Networks. *ICLR.* ICLR. Retrieved from https://arxiv.org/pdf/1511.06434.pdf

Reid, S. (2021, January). *https://www.scholarship.law.upenn.edu/.* Retrieved from The Deepfake Dilemma: Reconciling Privacy and First Amendment Protections: https://scholarship.law.upenn.edu/jcl/vol23/iss1/5/

Schreiner, M. (2022, April 28). *Deepfakes: How it all began - and where it could lead us*. Retrieved from https://the-decoder.com: https://the-decoder.com/history-of-deepfakes/

Segovia, K. Y., & Bailenson, J. N. (2009). Virtually true: children’s acquisition of false memories in virtual reality. *Media Psychology*, 371-393. Retrieved from https://stanfordvr.com/mm/2009/segovia-virtually-true.pdf

Stupp, C. (2019, August 30). *Fraudsters Used AI to Mimic CEO’s Voice in Unusual Cybercrime Case*. Retrieved from https://www.wsj.com/: https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402

Suwajanakorn, S., Seitz, S. M., & Shlizerman, I. K. (2017). *Synthesizing Obama: Learning lip sync from audio.* ACM Transactions on Graphics. Retrieved from https://grail.cs.washington.edu/projects/AudioToObama/siggraph17_obama.pdf

Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., & Niessner, M. (2016). Face2Face: Real-Time Face Capture and Reenactment of RGB Videos. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* Retrieved from https://openaccess.thecvf.com/content_cvpr_2016/html/Thies_Face2Face_Real-Time_Face_CVPR_2016_paper.html

Tremaine, D. W. (2019). *Two New California Laws Tackle Deepfake Videos in Politics and Porn*. Retrieved from www.dwt.com: https://www.dwt.com/insights/2019/10/california-deepfakes-law

Vaccari, C., & Chadwick, A. (2020). Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty and Trust in News. *Social Media + Society*, 1-13.

Virginia Legislative Information System. (2019). *Bill Tracking - 2019 session > Legislation* . Retrieved from https://lis.virginia.gov/: https://lis.virginia.gov/cgi-bin/legp604.exe?191+ful+CHAP0490&191+ful+CHAP0490

Westerlund, M. (2019). The Emergence of Deepfake Technology: A Review. *Technology Innovation Management Review*, 39-47.

WIPO. (2019, December 2019). Draft Issues Paper on Intellectual Property Policy and Artificial Intelligence. United States of America.

Zeman, E. (2021). Insurers Face Evolving Cyberrisk From Costly Hacks, Deepfake Attacks and Sophisticated Ransomware. *Best's Review*, 48-52.