Name of the author- Arpita Ananya Mohapatra

Year of study and programme enrolled- 5th year, BA LLB (Hons)

University- Madhusudan Law University, Odisha

Email Address- arpita.ananya00@gmail.com

Theme- Cyber Crime and Emerging Technologies

Sub-topic- Deepfake Technology

Title of the article- Unmasking Deepfake Menace: A Comprehensive Analysis of AI-Based Manipulation

Unmasking Deepfake Menace: A Comprehensive Analysis of AI-Based Manipulation

By:- Arpita Ananya Mohapatra

Unmasking Deepfake Menace: A Comprehensive Analysis of AI-Based Manipulation

Abstract

Deepfake technology has emerged as a concerning phenomenon in the digital age, enabling the creation of highly realistic and deceptive media through artificial intelligence techniques. This article explores the background and prevalence of deepfakes, delving into their design using advanced AI models such as Generative Adversarial Networks (GANs), Autoencoders, and First Order Motion Model. Notable deepfake incidents involving prominent figures like Barack Obama, Donald Trump, and Kim Jong-un are discussed, highlighting the potential real-world impacts.

This article explores the legal and ethical challenges posed by the rise of deepfakes. It delves into the legal frameworks surrounding deepfake technology, analyzes the ethical implications of their creation and use, examines existing responses, and proposes strategies for addressing their challenges. By reviewing the multidimensional aspects of deepfakes, this article seeks to contribute to the ongoing discourse on managing their impact in the digital age.

By adopting a multi-pronged approach and continually adapting to the evolving nature of deepfake technology, society can protect itself from the adverse effects of deepfakes and preserve the integrity of digital information and media content in the digital era.

KEYWORDS: Deepfake technology, AI-based manipulation, Privacy violations, Detection technology

I. Introduction

Deepfake technology, a portmanteau of "deep learning" and "fake," has rapidly emerged as a concerning phenomenon in the digital age. Deepfakes refer to manipulated digital media, often videos, that use artificial intelligence (AI) techniques to create highly realistic and deceptive content. These sophisticated manipulations involve replacing or superimposing someone's face onto another person's body, resulting in genuine and convincing videos.

Deepfake technology emerged in the 1990s through academic research and later gained popularity among online enthusiasts. A significant project known as the Video Rewrite program, published in 1997, modified the existing video footage of an individual speaking to depict that person mouthing the words in a different audio track.[1]

The Reddit community r/deepfakes played a significant role as the term 'deepfakes' was coined in 2017 by a Reddit user named "deepfakes".[2] As deepfakes gained traction, larger companies also adopted the technology to generate corporate training videos featuring deepfaked avatars and voices. Synthesia uses deepfake technology with avatars to generate customized videos is a prime example in this regard.[3] Moreover, the technology has the potential to resurrect the likeness of deceased individuals, further blurring the lines between reality and manipulated content.

---

[1] Bregler, Christoph; Covell, Michele; Slaney, Malcolm: "*Video Rewrite: Driving Visual Speech with Audio*", ACM Digital Library (13th of July, 2023, 10:53 AM), https://dl.acm.org/doi/10.1145/258734.258880

[2] Eyerys, A Reddit User Starts 'Deepfake' (last visited 12th of July, 2023)

[3] Chandler, Simon, "*Why Deepfakes Are A Net Positive For Humanity*", Forbes (13th of July, 2023, 6:34 PM), https://www.forbes.com/

II. Technology and Tools Used to Create Deepfakes

The most commonly used machine learning models to create deepfakes are:

i. Generative Adversarial Networks (GANs): GANs consist of two neutral networks; a generator and a discriminator. While the generator network creates synthetic data, such as a synthetic image that resembles the actual data provided in the output, the discriminator network assesses the authenticity of the synthetic data. It provides feedback to the generator on how to improve its output. After multiple repetitions, this process continues until the generator produces highly realistic synthetic data that is difficult to distinguish from the actual data.

ii. Autoencoders: Autoencoder is an unsupervised neural network capable of reducing the dimensionality of raw data and generating an output that replicates the input. Autoencoders comprise encoders and decoders. When data is input into the first layer, known as the input layer, of the autoencoder's neural network, the encoder compresses the data and forwards it to the decoder. The decoder's role is to reconstruct the original data. They help encode and decode facial features in a compressed form, making it easier to manipulate and create deepfake content.

Examples of notable deepfakes incidents

i. Barack Obama: On the 17th of April, 2018, a deepfake video of Barack Obama was posted on YouTube, where Barack Obama was seen cursing and calling Donald Trump names.[4] This video is intended to portray the horror of deepfakes.

ii. Donald Trump: On the 5th of May, 2019, a deepfake video of Donald Trump was posted on YouTube, based on a skit Jimmy Fallon performed on NBC's The Tonight Show.[5] In the original comedy sketch, Jimmy Fallon portrayed Donald Trump and engaged in a phone call with Barack Obama, where he humorously boasted about his victory in Indiana. In the deepfake version, Deepfakes transformed Jimmy Fallon's face to resemble Trump's face while keeping the audio unchanged to create humour and amusement.

iii. Kim Jong-un and Vladimir Putin: On the 29th of September, 2020, deepfakes of the North Korean leader Kim Jong-un and the Russian President Vladimir Putin were uploaded on YouTube, created by a nonpartisan advocacy group RepresentUs.[6] The deepfake videos featuring Kim and Putin were intended to be publicly aired as commercials, aiming to convey that interference by these leaders in US elections would have adverse consequences for American democracy.

iv. Pope Francis: In March 2023, an anonymous construction worker from Chicago used Midjourney to create a fake image of Pope Francis in a white Balenciaga puffer jacket, which went viral, receiving over twenty million views.[7]

---

[4] Fagan, Kaylee. "*A viral video that appeared to show Obama calling Trump a 'dips---' shows a disturbing new trend called deepfakes*", Business Insider India (the 12th of July, 2023, 4:10 PM), https://www.businessinsider.in/
[5] The Guardian, "The rise of the deepfake" (last visited 12th of July, 2023)
[6] MIT Technology Review, "Deepfake Putin" (last visited 17th of July, 2023)
[7] New York Post, "Pope Francis deepfake" (last visited 17th of July, 2023)

III. Legal Challenges Posed by Deepfakes

Online Defamation and Hate Speech: Deepfake videos falsely portraying individuals engaging in inappropriate or illegal activities can lead to defamation lawsuits and cause significant harm to a person's reputation. The use of deepfakes in spreading hate speech creates a toxic online environment.

Privacy Violations, Obscenity, and Pornography: Creating and disseminating deepfake videos without consent infringes upon an individual's right to privacy, leading to potential legal actions for invasion of privacy. Deepfakes are also used to create explicit or non-consensual content can result in legal action for revenge porn and violations of consent.

Intellectual Property Infringement: Deepfakes that utilize copyrighted materials, such as images or audio recordings, without proper authorization may lead to copyright infringement claims.

Fraud: Deepfake videos could be used to spread misinformation and interfere with elections, raising concerns about election integrity and voter manipulation. Deepfakes can be exploited for fraud, such as creating fake videos for blackmail, identity theft, or financial scams.

Cyberbullying and Harassment: Deepfake videos used for cyberbullying or harassment can have severe emotional and psychological effects on the victims, potentially leading to legal actions against the perpetrators.

Copyright and Fair Use: Legal questions arise concerning the fair use of deepfake technology for parody, criticism, or commentary versus copyright infringement when using copyrighted materials.

IV. Existing Legal Frameworks

Currently, India has no specific law or regulation to ban the use of deepfake technology. However, some specific statutes in India provide remedies against any offence committed by deepfake technology, albeit indirectly.

i. Constitution of India: Deepfakes violate the right to privacy, which is safeguarded by Article 21 of the Indian Constitution. The right to privacy has been held to be a fundamental right in the Puttaswamy case.[8].

ii. Information Technology Act, 2000: The IT Act is India's primary legislation governing electronic communications and digital transactions. Under Section 66 (computer-related offences) and Section 66-C (punishment for identity theft), crimes constituting identity theft and forgery can be prosecuted. Section 66D of the Act deals with situations where a communication device or computer resource is intentionally used for cheating or personation with malicious intent. Section 66E addresses privacy violations when an individual's images are captured, published, or transmitted in mass media without consent, infringing upon their privacy. Deepfakes containing pornographic content fall under Sections 67A and 66 B of the Act. Section 79 provides for the liability of intermediaries where the deepfake contents are posted.

[8] K.S. Puttaswamy & Anr v Union of India, (2017) 10 SCC 1

In the case of Myspace Inc. v Super Cassettes Industries Ltd[9], the Court held that in case of copyright infringement, notwithstanding the receipt of a Court order, the intermediaries are required to take down infringing content upon receiving a notification from the concerned private parties.

iii. Indian Penal Code, 1860: The IPC contains provisions that can be relevant to tackle deepfakes, such as Section 463 (Forgery), Section 464 (Making a false document), Section 465 (Punishment for forgery), and Section 469 (Forgery to harm reputation). Section 500 of the Act prescribes punishment for defamation, including online defamation.

iv. Copyright Act, 1957: Section 51 of the Act provides that the use of any property that belongs to another person on which the latter person has an exclusive right is violative of this Act. Section 57 gives the author the power to restrain or claim damages for any mutilation or distortion of his content that might damage the author's reputation.

In Amarnath Sehgal v Union of India[10]The Delhi High Court recognized the author's moral right over his work.

v. Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021: It was introduced to regulate digital content and intermediaries in India. Platforms hosting or distributing deepfake videos could be held accountable under these guidelines for not taking down harmful or illegal content.

vi. Personal Data Protection Bill, 2019: Deepfakes that involve the unauthorized use of an individual's personal information or likeness may fall under the purview of this law. This bill is designed to safeguard the personal data of individuals, encompassing information

[9] Myspace Inc. v Super Cassettes Industries Ltd, 236 (2017) DLT 478
[10] Amarnath Sehgal v. Union of India, 2005 (30) PTC 253 (Del)

about the concerned individual which has the capability of directly or indirectly identifying them.

V. Strategies for Addressing Deepfake Challenges

Advanced Detection Technology: Investing in research and development of sophisticated AI-based tools for detecting and identifying deepfake content can aid in quickly identifying and mitigating the spread of malicious deepfakes. Blockchain technology should be implemented, which allows individuals to track the source and alteration timeline of media, deterring the production and circulation of harmful deepfakes.

Legal Frameworks: Evaluating existing laws and regulations to determine whether they adequately address deepfake challenges and exploring the need for specific legislation targeting deepfakes is the need of the hour.

Platforms' Content Policies: Enforcing stricter content policies on social media and video-sharing platforms can help reduce the dissemination of harmful deepfake content and provide clear guidelines for addressing deep-fake-related issues.

Awareness Campaigns: Launching public awareness campaigns about the risks and implications of deepfakes can educate the public and encourage responsible consumption and sharing of digital content. Promoting media literacy and education initiatives can empower individuals to recognize potential deepfake manipulations.

International Cooperation: Promoting international cooperation and information-sharing among governments, law enforcement agencies, and technology companies can help address cross-border deepfake threats effectively.

VI. Conclusion

The rapid emergence of deepfake technology presents significant legal and ethical challenges in this age of technology. Deepfakes can be misused for defamation, privacy violations, fraud, and other harmful activities, leading to profound implications for individuals and society. Existing legal frameworks in India provide some remedies against deepfake offences indirectly. To tackle deepfake challenges, a multi-pronged approach is essential.

Moreover, international cooperation and information-sharing among governments and technology companies are crucial in combatting cross-border deepfake threats effectively. By adopting these strategies and continually adapting to the evolving nature of deepfake technology, society can better protect itself from the harmful impacts of deepfakes and preserve the integrity of digital information and media content in the digital era.