Proceedings of the 2nd International Conference on Global Politics and Socio-Humanities
DOI: 10.54254/2753-7048/77/2024.19110

Deepfake: Exploring the Societal Risks and Ethical Dilemmas of This New AI Technology

Tanghaorui Li[1,a,*]

[1]Culture Media and Creative Industries (Faculty of Arts & Humanities), King's College London, London, WC2R 2LS, United Kingdom
a. litanghaorui_rona@outlook.com
**corresponding author*

Abstract: On Social media platforms, fake news produced by AI has spread rapidly throughout these channels, negatively impacting millions of users. Deepfake, a form of AI communication that integrates deep learning, has become a hot topic at the communication theory and practice. This technology can produce artificial videos, images and sounds with high efficiency. While most research on deepfake focuses on detection systems in computer science, this paper argues that it needs to be analyzed from the perspectives of culture and communication. This paper uses literature review looking at how deepfake technology spreads false information, especially in influencing politics, invading people's privacy, and damaging trust in social media. By reviewing global studies and events, the research demonstrates that deepfake technology poses a significant danger to public opinion and personal privacy. The study also offers suggestions for addressing these issues, emphasizing the need for technological advancements and regulatory frameworks to guide future media and information management. Recommendations include continuous development of deepfake detection systems, legal constraints, and improving digital media literacy.

Keywords: deepfake, fake news, Artificial Intelligence.

1. Introduction

'Deepfake' is a hybrid term combining 'deep learning' and 'fake', first introduced in a report by the Center for Strategic and International Studies (CSIS) [1]. This technology commonly takes the form of face-swapping, where images or videos are merged, replaced, or superimposed to create AI-generated content that appears authentic [2]. Deepfake technology has not only impacted several fields but also raised serious ethical and social issues, such as the spread of fake news, political manipulation, and privacy violations.

This paper explores the societal risks posed by deepfake technology, as well as the ethical challenges it brings. Specifically, the paper explores how deepfake technology generates realistic fake videos and audio using artificial intelligence, leading to disinformation's widespread dissemination and its potential impact on the political, social and cultural spheres. The main methodology in this paper is literature review. By integrating global news events, this study offers a critical perspective on the far-reaching effects of deepfake technology on audiences. The research helps raise awareness of the potential threats of deepfake technology, especially in disseminating false information, political

DOI: 10.54254/2753-7048/77/2024.19110

manipulation, and invasion of personal privacy. In addition, the research provides policy and technological solutions for mitigating the risks posed by this technology.

2. How has Deepfake gradually become a popular topic?

In the Western context, deepfake technology came into public sight in 2017 when a Reddit user utilized the deepfake technique to change the face of a female protagonist in a pornographic video with that of a well-known actress [3]. This incident sparked widespread discussion, with users across the internet rapidly sharing this false content. Another example of an application is from China. In September 2019, ZAO, a product developed by Stranger Company, became an overnight sensation and ranked among the top free app downloads from various app stores in less than 24 hours. Users who uploaded a photo on ZAO were able to generate videos. However, the app was quickly pulled by authorities due to concerns about violations of portrait rights and privacy. The disruptive potential of deepfake lies in the technology's ability to be widely accessible—as almost anyone with a computer can produce fake videos that are virtually indistinguishable from real media [4].

Initially, deepfake research focused onactresses, political leaders, actresses, comedians, and entertainers whose faces were compiled in pornographic videos [5]. However, Maras and Alexandrou predict that future occurrences of deepfakes may be revenge porn, cyberbullying, false video evidence in court, political sabotage, terrorist propaganda, extortion, market manipulation, and fake news [2]. Deepfake technology can produce humorous, pornographic, or political content, portraying individuals saying or doing things without their consent [4].

3. Analysis: The threat of Deepfake

3.1. Difficult to Identify the Veracity of Information

Psychological and communication research widely agrees that people are often motivated to support and accept information consistent with their pre-existing views and beliefs [6]. Shu and her research partners illustrate an interesting correlation between the spread of fake news and psychological and cognitive theories. According to their findings, people are naturally poor at distinguishing between true and false information, especially when it confirms their ideological beliefs [7].

Unlike traditional fake news, which involves post-processed with images through tools such as Photoshop, deepfake is directly manipulated and generated through machine learning and artificial intelligence techniques to produce more deceptive visual and audio content. Due to the high degree of naturalness of the content it generates, it is often difficult for an average user to recognize the falsified elements.

For instance, in April 2023, U.S. President Joe Biden officially announced his decision to run for re-election. In response, the Republican National Committee released an entirely AI-generated video advertisement [8]. This video depicted a utopian world with imagined disasters if Biden won the re-election. Many audiences failed in identifying reality because of the realistic graphics and images. As a result, many people were misled and influenced by the fabricated content. This case demonstrates that deepfake technologies not only can easily confuse the general public but also pose a serious threat to global political, economic, social, and cultural stability. As these technologies continue to evolve rapidly, their potential to undermine public trust and spread disinformation on a global scale becomes increasingly concerning.

3.2. High Communication Efficiency of Mis/Disinformation

While natural language processing and generative techniques, at the heart of generative AI technologies, have evolved into complex, large language models (LLM) capable of reading, writing,

DOI: 10.54254/2753-7048/77/2024.19110

and interpreting text, these models excel at mimicking online discourse. In addition to grammatical rules, they are prone to reproducing human biases and exacerbating the dissemination of disinformation.

Peter's concept of 'Algorithmic Political Bias' illustrates how cognitive biases can be a tool for AI to present users constantly seeing their opinionated perspectives [9].

In addition, research from the University of Zurich shows that misinformation generated through automated methods is often more convincing than false content produced by humans [10]. This is because AI-generated content can reduce the cost of producing misinformation while amplifying its sensationalism and emotional impact. The spread of false information online is very likely to provoke emotional reactions, reinforce individual biases, and disrupt the process of forming social consensus.

In other words, audiences are inclined to accept it as truth without critical consciousness when deepfake news aligns with their pre-existing beliefs. As social media platforms increasingly prioritize content that resonates emotionally, the audience become more susceptible to misinformation. Deepfake contents, therefore, may strengthen the spread of misinformation because they weaken trust in reliable media sources.

3.3. Low efficiency in information verifying process

The proliferation of AI language models has made generating misinformation more cost-effective and widespread. Moreover, the massive work load generated by AI social media bots makes it harder for fact-checkers to keep up with the content volume and increases the difficulty of distinguishing between real and fake information. For example, during the Russia-Ukraine war, numerous social media bots have been responsible for generating and disseminating disinformation.

Moreover, the interactivity of AI chatbots enables them to create highly personalized, targeted narratives. The seamless blending of AI-generated media with accurate content blurs the lines between truth and falsehood, further complicating the verification process [11]. Timeliness also affects the experience of social media users. A study of 2,005 respondents revealed that, while not all were entirely deceived by political deepfakes, many expressed uncertainty about the authenticity of the content they encountered. This uncertainty undermines public trust in media platforms because readers become less confident distinguishing between real and fake news [12]. Scholar Temir also notes that this increasing mistrust exacerbates the challenges faced by fact-checkers, as even credible sources are met with skepticism [13].

4. How do we respond and protect?

On the technicla front, there is a need to develop detection software and improve the technical ability to distinguish between fake and real contents. In the Internet era, social media platforms provide a viral pathway for disseminating deeply falsified information. Therefore, these platforms must take responsibility for regulating and censoring false content while ensuring the protection of individuals' rights. First, platforms should establish the legitimacy for using technology and prohibit the circulation of unauthorized, deeply falsified information on the platform. For example, on March 21st, 2023, TikTok updated the platform's community guidelines to prohibit the production of deepfake videos featuring personalities other than public figuresand any deepfake videos depicting real-life scenarios must indicate that it has been synthesized or altered. Second, platforms should invest in AI-powered detection systems capable of high-frequency, round-the-clock monitoring to enhance content regulation [14].

AI-related legislation can also contribute to the construction and realization of social justice at a time when digital technologies are widely used. Unregulated AI usage can exacerbate social injustice and produce unfair outcomes. A transparent and fair AI system requires a legal framework to prevent

DOI: 10.54254/2753-7048/77/2024.19110

abuse and bias and ensuring that the benefits of AI are distributed equitably. In this context, legal measures can protect the authenticity of information by clearly defining the boundaries of deepfake usage.

Several countries have successfully enacted laws and regulations on deepfake technology to regulate its development and standardize the border of deepfake use.

For instance, Since 2012, The United States has established a relatively comprehensive regulatory framework for AI, and federal laws have been passed to continuously improve oversight in this area. [15]

Similarly, Chinese laws on Deepfake also focus on protecting portrait rights, personal information, and national security. In July 2023, China's Cyberspace Administration of China (CAC) and six other regulators promulgated the 'Interim Measures for the Administration of Generative Artificial Intelligence Services', explicitly supporting innovation while categorizing and grading AI services to ensure compliance with privacy and security standards [16].

Deep forgery technology has become increasingly mature and promising, and there is an urgent need for newer laws to regulate its use in the marketplace. Deepfake technology is generated, as Mark Warner (who wrote a white paper on social media regulation) describes as "unprecedented wave of false or defamatory content" [17]. Addressing the issue caused by deepfake is becoming increasingly critical as it is the most recent deception in online video communication. To mitigate its effects, it is essential to enhance digital media literacy among audiences, thereby improving their ability to recognize false information. Given the high degree of authenticity, ubiquity and rapid evolution of in-depth forgery technology, it is necessary for audiences to continuously improve their personal media literacy, cultivate the ability to critically analyze media contents and deepen their understanding and ability to recognize deepfake technology.

Understanding the rules of social platforms is the first step. Audiences should gain a foundational knowledge of the technology which can help them to be more vigilant when encountering suspicious content and prevent them from being easily deceived or spreading false information. At the same time, digital media literacy requires developing critical thinking and judgment. Audiences with higher levels of digital media literacy will be more adept at to identifying disinformation and protecting their rights in the digital environment. Through education and training, audiences can learn how to verify sources, check facts, and recognize false or manipulated information. Readers need to learn how to objectively assess the truthfulness and reliability of information and be as unaffected as possible by emotional terminology and personal bias. For example, when confronted with material that deviates from their point of view, audiences should remain open-minded and seek information from multiple sources to avoid falling into the 'echo chamber' [18].

5. Discussion and Conclusion

While generative AI brings about groundbreaking innovations, it also raises ethical dilemmas, particularly concerning data bias and algorithmic bias. As AI systems rely heavily on large datasets for training, there is a risk that these systems will inadvertently learn and amplify existing biases, leading to unfair or discriminatory outcomes in decision-making processes. Deepfake technology, a byproduct of generative AI, has become a popular research topic across academic disciplines and industries alike. Its potential to cause social and privacy-related harm through the spread of disinformation poses a serious threat.

Current AI capabilities for detecting false content lag behind their abilities to generate it. In the post-truth era, legal frameworks alone will not suffice to protect individuals from the dangers of disinformation. The public themselves also need to improve their digital media literacy and become critical readers.

DOI: 10.54254/2753-7048/77/2024.19110

References

[1] Lewis, James Andrew, and Arthur Nelson. "Trust Your Eyes? Deepfakes Policy Brief." CSIS, www.csis.org/analysis/trust-your-eyes-deepfakes-policy-brief. Accessed 17 Oct. 2024.

[2] Maras, Marie-Helen, and Alex Alexandrou. "Determining Authenticity of Video Evidence in the Age of Artificial Intelligence and in the Wake of Deepfake Videos." The International Journal of Evidence & Proof, vol. 23, no. 3, 28 Oct. 2018, pp. 255–262, doi:10.1177/1365712718807226.

[3] Cole, Samantha, et al. "Ai-Assisted Fake Porn Is Here and We're All Fucked." VICE, 9 Aug. 2024, www.vice.com/en/article/gal-gadot-fake-ai-porn/. Accessed 17 Oct. 2024.

[4] Fletcher, John. "Deepfakes, Artificial Intelligence, and Some Kind of Dystopia: The New Faces of Online Post-Fact Performance." Theatre Journal, vol. 70, no. 4, 2018, pp. 455–471, doi:10.1353/tj.2018.0097.

[5] Maras, Marie Helen, and Alex Alexandrou. "A Study on Deep Fake Face Detection Techniques." Sage Journals, vol. 23, no. 3, doi.org/10.1177/13657127188072.

[6] Edgerly, Stephanie, et al. "When Do Audiences Verify? How Perceptions about Message and Source Influence Audience Verification of News Headlines." Journalism & Mass Communication Quarterly, vol. 97, no. 1, 5 Aug. 2019, pp. 52–71, doi:10.1177/1077699019864680.

[7] Shu, Kai, et al. "Fake News Detection on Social Media." ACM SIGKDD Explorations Newsletter, vol. 19, no. 1, Sept. 2017, pp. 22–36, doi:10.1145/3137597.3137600.

[8] RNC Counters Biden Announcement with Dystopian, AI-Aided Video - The Washington Post, www.washingtonpost.com/politics/2023/04/25/rnc-biden-ad-ai/. Accessed 17 Oct. 2024.

[9] Peters, Uwe. "Algorithmic Political Bias in Artificial Intelligence Systems." Philosophy & Technology, vol. 35, no. 2, 30 Mar. 2022, doi:10.1007/s13347-022-00512-8.

[10] Spitale, Giovanni, et al. "Ai Model GPT-3 (Dis)Informs Us Better than Humans." Science Advances, vol. 9, no. 26, 30 June 2023, doi:10.1126/sciadv.adh1850.

[11] reuters_tickers. "Russia Using Generative AI to Ramp up Disinformation, Says Ukraine Minister." SWI Swissinfo.Ch, Www.Swissinfo.Ch, 16 Oct. 2024, www.swissinfo.ch/eng/russia-using-generative-ai-to-ramp-up-disinformation, -says-ukraine-minister/87742674. Accessed 17 Oct. 2024.

[12] Vaccari, Cristian, and Andrew Chadwick. "Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News." Social Media + Society, vol. 6, no. 1, Jan. 2020, doi:10.1177/2056305120903408.

[13] Temir, Erkam. "Deepfake: New Era in The Age of Disinformation & End of Reliable Journalism." Journal of Selcuk Communication, vol. 13, no. 2, 2020.

[14] Vincent, James. "TikTok Bans Deepfakes of Nonpublic Figures and Fake Endorsements in Rule Refresh." The Verge, 21 Mar. 2023, www.theverge.com/2023/3/21/23648099/tiktok-content-moderation-rules-deepfakes-ai. Accessed 17 Oct. 2024.

[15] Langa, Jack. "Deepfakes, Real Consequences: Crafting Legislation to Combat Threats Posed by Deepfakes." Boston University Law Review, vol. 101, no. 2, March 2021, pp. 761-802. HeinOnline, https://heinonline.org/HOL/P?h=hein.journals/bulr101&i=767.

[16] Wong, Daisy, et al. "China: New Measures on Generative Artificial Intelligence." DLA Piper, www.dlapiper.com/en-us/insights/publications/2023/07/china-new-measures-on-generative-artificial-intelligence. Accessed 17 Oct. 2024.

[17] Warner, Mark. Potential Policy Proposal for Regulation of Social Media and Technology Firms, 2018.

[18] He, Yiqing, et al. "Information Cocoons on Short Video Platforms and Its Influence on Depression among the Elderly: A Moderated Mediation Model." Psychology Research and Behavior Management, Volume 16, July 2023, pp. 2469–2480, doi:10.2147/prbm.s415832.