

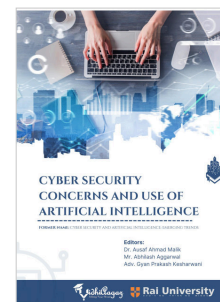
CHAPTER 37

A Critical Study about Combating the Dark Side of AI: Deepfake Prevention through Ethical AI Development

Ankita Das and Manish Kumar Singh

in *Cyber Security Concerns and Use of Artificial Intelligence*, edited by Ausaf Ahmad Malik, Abhilash Aggarwal and Gyan Prakash Kesharwani (VidhiAagaz 2024), pp. 308–314

ISBN 978-81-947620-5-8



CITATIONS

BLUEBOOK (21ST ED.)

Ankita Das & Manish Kumar Singh, *A Critical Study about Combating the Dark Side of AI: Deepfake Prevention through Ethical AI Development*, in *CYBER SECURITY CONCERNS AND USE OF ARTIFICIAL INTELLIGENCE* 308 (Ausaf Ahmad Malik, Abhilash Aggarwal & Gyan Prakash Kesharwani eds., VidhiAagaz 2024).

OSCOLA (4TH ED.)

Ankita Das and Manish Kumar Singh, 'A Critical Study about Combating the Dark Side of AI: Deepfake Prevention through Ethical AI Development' in Ausaf Ahmad Malik, Abhilash Aggarwal and Gyan Prakash Kesharwani (eds), *Cyber Security Concerns and Use of Artificial Intelligence* (VidhiAagaz 2024) 308.

CHICAGO (17TH ED.)

Das, Ankita, and Manish Kumar Singh. "A Critical Study about Combating the Dark Side of AI: Deepfake Prevention through Ethical AI Development." In *Cyber Security Concerns and Use of Artificial Intelligence*, edited by Ausaf Ahmad Malik, Abhilash Aggarwal and Gyan Prakash Kesharwani, 308–314. Maihar: VidhiAagaz, 2024. <https://vidhiaagaz.com/books/cyber-security-concerns-and-use-of-artificial-intelligence/chapter-37/>.

APA (7TH ED.)

Das, A., & Singh, M. K. (2024). A critical study about combating the dark side of AI: Deepfake prevention through ethical AI development. In A. A. Malik & A. Aggarwal & G. P. Kesharwani (Eds.), *Cyber security concerns and use of artificial intelligence* (pp. 308–314). VidhiAagaz. <https://vidhiaagaz.com/books/cyber-security-concerns-and-use-of-artificial-intelligence/chapter-37/>

SOURCE

<https://vidhiaagaz.com/books/cyber-security-concerns-and-use-of-artificial-intelligence/chapter-37/>

RIGHTS

Free to read online. © VidhiAagaz 2024. All rights reserved: no part of this publication may be copied, reproduced, adapted or transmitted in any form or by any means without the prior written permission of the publisher.

Citations are given as a guide; check them against the style manual you follow. This page was added by VidhiAagaz to the downloaded file. The chapter follows, with the page numbers of the printed book.

A CRITICAL STUDY ABOUT COMBATING THE DARK SIDE OF AI: DEEPFAKE PREVENTION THROUGH ETHICAL AI DEVELOPMENT

Ankita Das⁵⁵ and Dr Manish Kumar Singh⁵⁶

ABSTRACT

Deepfake technology has changed the way cybercrime is done and brought up important legal and ethical issues. This Article talks about how deep-fakes hurt weak groups, mainly by being used for bad things like spreading false information about politics, sharing sexual content without permission, and stealing identities. Artificial intelligence makes movies and pictures that look very real but are actually fake. Many people, especially women and famous people, lose money, their name, and their emotions because of deep-fakes. As deepfakes keep spreading, many are afraid that they could hurt trust in digital media and information that is open to the public. This Article is critical study of the ethical issues related to deep-fakes, with a focus on permission, privacy, and accountability. These problems with different countries' laws are brought to light because the current legal systems aren't thought to be strong enough to handle these issues. This Article looks at recent changes in the law and case studies from around the world to explore real-world answers to the issues raised by deepfakes. We need strong rules and standards that involve governments, IT companies, and regular people working together. At the end of the Research, it is suggested that people from different countries work together to combat the moral issues raised by deepfakes, while also encouraging digital responsibility and protecting those who are most at risk.

Keywords: Deep-fake technology, Cybercrime, Ethical challenges, Artificial intelligence (AI), Non-consensual pornography.

INTRODUCTION

Deepfake technology is maybe the most intriguing and concerning AI advancement. Deepfakes use AI algorithms, specifically generative adversarial networks, to create lifelike yet fake media. These technologies enable speech manipulation, visually engaging content that is practically indecipherable from actual media, and video face swapping. Deepfakes were first popular in gaming and entertainment, but their possible destructive applications have raised digital age worries.⁵⁷ Artificial intelligence and machine learning have made deepfakes popular. Since deepfake production tools become more available, they have been used in online communities and social media more and more. What started as a niche app has become a global phenomenon with major ramifications, especially in disinformation, identity theft, and unauthorised sharing. This article examines the social, legal, and ethical ramifications of deepfakes to promote accountability and prevention. This essay examines the dark side of deepfake technology and its threats to digital integrity, consent, and privacy. It will raise awareness of deepfake usage, show the need for tighter legislation, and show tech companies, social media platforms, and AI

⁵⁵ Research Scholar at NIMS school of Law (NIMS University Rajasthan), Jaipur, India.

⁵⁶ Assistant Professor at NIMS school of Law (NIMS University Rajasthan), Jaipur, India.

⁵⁷ Seng, L. K., Mamat, N., Abas, H., & Ali, W. N. H. W. (2024). AI Integrity Solutions for Deepfake Identification and Prevention. Open International Journal of Informatics, 12(1), 35-46.

developers' moral responsibilities to solve it. This essay will evaluate deepfakes' ethical issues and suggest answers and collaborations.

UNDERSTANDING DEEPPFAKES: TECHNOLOGY AND APPLICATIONS

Artificial intelligence subfield generative adversarial networks (GANs) build deepfakes. GANs have opposing generator and discriminator neural networks. The generator creates fake images and videos, and the discriminator compares them to real data to determine realism. This confrontational method helps the generator develop hyper-realistic fake news with practice. These AI models employ large datasets like photographs or videos of a person to mimic their voice or appearance.⁵⁸ Despite their bad image, deepfakes have several legitimate uses. Entertainment filmmakers can de-age or recreate actors using deepfakes for visual effects. But deepfakes' dark side is troubling. One of the worst uses of deepfakes is to spread political disinformation by editing politician or other public figure videos. By superimposing women's faces over graphic material without their consent, non-consensual pornography causes considerable psychological and emotional suffering.

ETHICAL CHALLENGES OF DEEPPFAKE TECHNOLOGY

Deepfake technology raises ethical concerns since it may violate consent and privacy. Non-consensual pornography in deepfake stands out. Sexually graphic images are often painted on people's faces without their consent, causing emotional, psychological, and social harm. Digital identity manipulation victims sometimes feel helpless since the legal system is unprepared to address such complicated situations. Deepfakes can impersonate people and threaten their privacy and security, in addition to non-consensual pornography.⁵⁹ Deepfake pornography disproportionately impacts women and other marginalised individuals and reinforces harmful stereotypes and intolerance. Movies and images are often manipulated to spread misinformation, harm reputations, and hurt leaders and celebrities. Deepfake creation tools are readily available, compounding these issues. Anyone with the right tools can spread dangerous information, mostly to conspicuous or weak targets.

Deepfakes exacerbate misinformation and online media distrust. Political disinformation campaigns are increasingly using deepfakes to sway public opinion. Many platforms can't moderate all content⁶⁰, and their procedures vary. Last but not least, those who create and propagate harmful deepfakes are culpable, but the internet and worldwide legal issues make it hard to hold them accountable. Lack of transparency increases the potential of deepfakes causing substantial damage.

LEGAL FRAMEWORKS: GLOBAL CHALLENGES AND GAPS

Deepfake technology is threatening legal systems worldwide. Deepfake laws are struggling to keep up with their rapid development. Some countries have been ahead of the curve in regulating deepfake

⁵⁸ Widder, D. G., Nafus, D., Dabbish, L., & Herbsleb, J. (2022, June). Limits and possibilities for "Ethical AI" in open source: A study of deepfakes. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 2035-2046).

⁵⁹ Romero Moreno, F. (2024). Generative AI and deepfakes: a human rights approach to tackling harmful content. *International Review of Law, Computers & Technology*, 1-30.

⁶⁰ Whyte, C. (2020). Deepfake news: AI-enabled disinformation as a multi-level public policy challenge. *Journal of cyber policy*, 5(2), 199-217.

usage, while others have lagged, leaving gaps that criminals can exploit.

Current Legal Frameworks

Certain countries have limited the harmful use of deepfakes in recent years. The EU's General Data Protection Regulation (GDPR) protects individuals' right to control their personal data, including likeness and photos. GDPR allows individuals to seek the removal of deepfakes that violate their privacy rights, and violators can be substantially fined. GDPR addresses privacy issues, although deepfakes are not directly addressed. Few US jurisdictions have passed laws against deepfake and other non-consensual pornography.⁶¹ New laws in Texas and California make deepfakes for political campaigns and non-consensual pornography unlawful. AB 602 allows California victims of deepfake pornography to sue for damages. Deepfakes may also be susceptible to UK revenge pornography laws. Because there is no law addressing non-consensual deepfakes, victims have little options. South Korea, along with other Asian nations, has cracked down on cybersex crimes that use deepfake technology. These laws criminalise deepfake pornography production and distribution with serious penalties.

Legal Gaps

Despite progress, deepfake regulation worldwide is still lacking. Thus, victims are at risk in many countries without deepfake laws.⁶² Since deepfake material can be generated in one country and shared worldwide, the internet's convoluted jurisdictional structure makes prosecution harder. Because deepfakes involve privacy, data protection, and intellectual property, present policies often fail to address them. Many countries still utilise outdated copyright, defamation, and fraud laws to combat deepfake usage. Internet anonymity makes deepfake prosecution harder. Even when dangerous content is identified, encrypted platforms or servers in locations with weak prohibitions make it difficult to trace its source. Deepfake victims must prove the content injured them. This is especially difficult with political influence and identity theft, where the harm is hard to measure but long-lasting.

Case Studies

Several major cases highlight the challenges court systems have in addressing deepfakes. In 2019, a deepfake video of the Belgian prime minister giving an insane climate change address went viral, causing great confusion until it was debunked. Despite no legal penalties, the film showed how deepfakes can disrupt politics. In another high-profile UK case, a woman's face was clearly featured on sexually graphic recordings. Despite her best efforts, she could not legally remove the recordings, and the perpetrator was never caught. This scenario showed why deepfake abuse laws are needed. A viral deepfake video misrepresented a political candidate in the 2020 Delhi elections in India, eliciting intense public anger. The country's outdated misinformation laws made it difficult to punish digital criminals, demonstrating the court system's inability to handle emerging crimes.

Cross-Border Complexities

⁶¹ Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology innovation management review*, 9(11).

⁶² Labrecque, L. I., Peña, P. Y., Leonard, H., & Leger, R. (2024). Not all sunshine and rainbows: exploring the dark side of AI in interactive marketing. *Journal of Research in Interactive Marketing*.

Controlling deepfakes is difficult because the internet is available worldwide. You can make deepfake content in one country, post it to another, and see it elsewhere. In nations with weak or nonexistent deepfake legislation, offenders might exploit legal loopholes, making enforcement difficult owing to cross-border complexity. If a deepfake developer is based in a country without digital content modification laws, they are practically impossible to prosecute. Global collaboration is crucial, but cybercrime and intellectual property regimes often fail to address deepfake technologies.⁶³ Governments, tech companies, and international bodies must work together to regulate deepfakes. International treaties and regulations on deepfake technology exploitation should be adopted to handle these digital risks and overcome jurisdictional limitations. Deepfakes will exploit legal loopholes to prey on individuals and organisations till then.

ETHICAL AI DEVELOPMENT: PREVENTIVE MEASURES AGAINST DEEPFAKES

AI's rapid development of hyper-realistic deepfakes creates serious moral issues. Politicians, tech corporations, and AI experts must develop ethical methods to fight deepfakes. These challenges must be resolved to preserve individual rights, stop harmful content, and ethically develop AI. AI developers are crucial to fighting deepfake technology. Starting with ethics in AI system design is critical. In deepfake development and detection algorithms, developers must take measures. Deepfakes use GANs to simulate human behaviour. These algorithms must not be abused. AI should detect and report deepfake content to reduce dangers. Engineers should put ethics into AI systems from the start to prevent misuse. AI models' societal impact can be assessed using ethical auditing. Malicious deepfakes damage women and marginalised groups more than men. Software programmers can fix this by making datasets more representative of society and preventing AI systems from biasing data.

Role of IT Companies

IT, social media, and AI developers must stop deepfakes. Facebook, Twitter, and TikTok need tougher content monitoring since they distribute deepfake content. Tech firms may invest in deepfake detection. While many social media sites use AI-based algorithms to delete unwanted content, advanced technologies are needed to detect deepfakes before they spread virally. Platform users should check material.⁶⁴ Label altered or questionable media for viewer choice. Technology businesses should enforce community norms and punish deepfake content authors. Clear penalties like permanent bans or suspensions may deter repeat offenders. IT firms that boost privacy settings and give users more digital identity control limit deepfake manipulation.

AI and Human Rights

Human rights concerns surround AI-created deepfakes. Deepfake technology endangers dignity, autonomy, and security. Non-consensual pornography and political disinformation degrade people by

⁶³ Tremont, T. M. (2023). *Human-AI: Using Threat Intelligence to Expose Deepfakes and the Exploitation of Psychology* (Doctoral dissertation, Capitol Technology University).

⁶⁴ Fletcher, J. (2018). Deepfakes, artificial intelligence, and some kind of dystopia: The new faces of online post-fact performance. *Theatre Journal*, 70(4), 455-471.

distorting their images. Other ethical issues exist beyond autonomy.⁶⁵ Fakes cause identity loss by making people say and do things they never did. Identity manipulation can have long-term implications on victims' personal and professional lives. Deepfake attacks on celebrities can alter public opinion and harm reputations. Deepfakes can also swindle, harass, and extort people, endangering their security. The complexity of AI-generated video makes deepfakes harder to spot, making them easier for bad actors to use. Legislators and AI developers must consider human rights while regulating deepfakes.

Emerging Technologies for Detection

Digital technologies to detect and stop deepfakes are needed as they get more complex. Because of this, AI-powered detection solutions have evolved. Deeptrace and Truepic are creating digital content change detection tools. Video and photo AI may detect unnatural face gestures, lighting, and audio-visual sync issues. Blockchain⁶⁶ and AI-based detection may assist verify digital content authenticity, which is growing more important. An immutable blockchain ledger preserves every file's origin and modification history, giving digital content makers confidence. The government may mandate media, law enforcement, and political campaigns use these technology to promote their adoption. Ethical AI research, proactive IT controls, and better detection tools are needed to mitigate AI-driven deepfakes. Ethical AI development and industry collaboration can stop deepfakes and restore digital media confidence.

THE ROLE OF SOCIAL MEDIA PLATFORMS

Social media helps spread and stop deepfakes. These platforms can prohibit deepfakes because they disseminate most digital material globally. To prevent deepfakes, platforms must manage content, reduce harm, and work with authorities and governments. Social media must moderate. Deepfakes transmit deceptive information, non-consensual pornography, and change users' identities. Social media, video sharing, and app firms are under pressure to stop them. Preventing deepfake content harm requires proactive content management. To stop its spread, promptly identify, mark, and remove modified media.

Digital material anomalies including distorted voice patterns, unusual lighting, and strange face motions can help automatic AI recognise deepfakes. AI alone won't work as deepfakes improve. Social networks need AI detection and human monitoring. By evaluating flagged content, skilled moderators can quickly eliminate dangerous content while keeping genuine submissions. Moderators must also be open. Platforms must explain how and why they delete deepfake content. Transparency about content removal and violation criteria creates trust and accountability. Social media platforms should allow users to report suspicious information to combat deepfakes.

Many social media companies tackled deepfake harm. Facebook prohibits election-influencing modified media. The company has enhanced real-time deepfake detection algorithms through academic cooperation and research. Facebook's 2020 "Deepfake Detection Challenge" trained AI to detect

⁶⁵ Qureshi, J., & Khan, S. (2024). Artificial Intelligence (AI) Deepfakes in Healthcare Systems: A Double-Edged Sword? Balancing Opportunities and Navigating Risks.

⁶⁶ Godulla, A., Hoffmann, C. P., & Seibert, D. (2021). Dealing with deepfakes—an interdisciplinary examination of the state of research and implications for communication studies. *SCM Studies in Communication and Media*, 10(1), 72-96.

deepfakes. Twitter now flags and removes misleading or harmful content. This includes scary or untrue deepfake videos.⁶⁷ Twitter tags help users evaluate content. YouTube, another big site, has also addressed deepfakes and political candidate falsehoods. Platforms will eliminate risky deepfake films. Films with non-consensual pornography, harassment, and fraud are included. Despite these efforts, many believe platforms should do more to stop deepfakes due to their prevalence.

Social media platforms realise they cannot fight deepfakes alone. Legal frameworks must be developed with governments for digital accountability and transparency. Governments may ban deepfakes on platforms. The GDPR prioritises privacy, but the DSA requires social media transparency and responsibility. Multinational task forces can share intelligence, confront new threats, and coordinate enforcement. Finally, social media must limit deepfake influence. Platforms can reduce deepfake content and make the internet safer by improving content control, harm mitigation, and government regulation. Companies may protect digital media ethics by being honest and proactive.

PROTECTING VULNERABLE GROUPS: GENDER-BASED VIOLENCE AND DISCRIMINATION

Women are being disproportionately targeted by non-consensual pornography and digital manipulation due to deepfake technology. Deepfake abuse victims, who are mostly women and marginalised people, often suffer psychological, social, and financial harm from having their faces on sexually explicit images and films without their consent. This exploitation reinforces a culture that regards women as objects by marginalising women and encouraging harmful gender norms. Race and LGBTQ+ people, who already face systemic prejudice, may be further victimised by deepfakes. People with bad motives are more likely to manipulate media to shame, blackmail, or degrade these communities. Deepfakes make it easy to create and disseminate fake news without penalties, worsening social inequality. Gender-based assault victims of deepfakes get legal and emotional support from multiple global groups. The UK has banned "revenge porn," which has helped prosecute deepfakes and other non-consensual pornography cases. Deepfake misuse might also trigger GDPR privacy laws.⁶⁸ However, these measures are still disconnected and require stronger, more comprehensive legislative frameworks and international collaboration to protect vulnerable groups. Governments, tech corporations, and advocacy groups must work together to improve detection, laws, and safe spaces for deepfake-related gender-based violence and prejudice victims.

TOWARDS DIGITAL RESPONSIBILITY: RECOMMENDATIONS FOR ETHICAL USE OF AI

We must fight deepfakes and establish AI ethics soon and collaboratively. Governments, tech corporations, and civil society must collaborate to pass legislation that promote ethical AI use and ban deepfakes. Deepfake technology abuse must be a cybercrime with sanctions by governments. To

⁶⁷ Battista, D. (2024). Political communication in the age of artificial intelligence: an overview of deepfakes and their implications. *Society Register*, 8(2), 7-24.

⁶⁸ Samuel-Okon, A. D., Akinola, O. I., Olaniyi, O. O., Olateju, O. O., & Ajayi, S. A. (2024). Assessing the Effectiveness of Network Security Tools in Mitigating the Impact of Deepfakes AI on Public Trust in Media. *Archives of Current Research International*, 24(6), 355-375.

prevent dangerous deepfakes, AI engineers must integrate ethical checks. To stop the spread of altered material, governments and tech companies should create universal content moderation norms, which would require more internet platform monitoring. Public education and awareness are key to defeating deepfakes.⁶⁹ Digital literacy programmes should warn about deepfake technology and teach fake news detection and reporting. Teaching people to evaluate digital content's legitimacy can lessen deepfakes' social impact, deception and exploitation. For an informed and resilient populace, educational institutions, social media sites, and news organisations must participate in these initiatives.

Internet and deepfake threats are worldwide, requiring swift international coordination. Countries should form partnerships to address deepfake morality and exchange detection science. Hold criminals accountable International coalitions fighting cybercrime and data privacy could exploit jurisdictional loopholes. The combined effort would reduce global deepfake damage and boost digital responsibility. AI developments are needed for technology, but ethics must be balanced to minimise harm. AI research and development should be ethical to avoid societal injustice and human rights breaches. To guarantee AI developments are thoroughly examined before implementation, governments and tech corporations should create ethical regulatory agencies or advisory boards.⁷⁰ Tech companies can gain responsibility and confidence in their AI systems by revealing how their algorithms detect and handle deepfakes. Ethics allow us to investigate AI's potential without endangering society or human security. Finally, governments, industry leaders, and the public must handle deepfakes holistically. Implementing detailed policy frameworks, expanding digital literacy, boosting international collaboration, and preserving ethical AI development may balance technology innovation and social responsibility. Digital media integrity and online predator protection are key to fighting deepfakes beyond tech regulation.

CONCLUSION

In conclusion, this article has examined the ethical, legal, and societal implications of deepfake technology, highlighting the urgent need for robust regulatory frameworks and ethical AI development. We emphasized the disproportionate impact of deepfakes on vulnerable groups and the necessity for proactive measures from governments, tech companies, and civil society. A global response is essential to combat the challenges posed by deepfakes effectively. Looking ahead, fostering ethical AI development presents an opportunity to harness the positive potential of artificial intelligence as a force for good, promoting transparency, accountability, and digital responsibility in an increasingly complex digital landscape.

⁶⁹ Pantserov, K. A. (2020). The malicious use of AI-based deepfake technology as the new threat to psychological security and political stability. *Cyber defence in the age of AI, smart societies and augmented humanity*, 37-55.

⁷⁰ Wagner, T. L., & Blewer, A. (2019). "The word real is no longer real": Deepfakes, gender, and the challenges of ai-altered video. *Open Information Science*, 3(1), 32-46.