

Deepfakes, Misinformation, and Disinformation in the Era of Frontier AI, Generative AI, and Large AI Models

Mohamed R. Shoaib, Zefan Wang, Milad Taleby Ahvanooy, Jun Zhao
School of Computer Science and Engineering
Nanyang Technological University
Singapore

All authors contribute equally to the paper.

Mohamedr003@e.ntu.edu.sg, Zefan.wang@ntu.edu.sg, Milad.ta@ntu.edu.sg, Junzhao@ntu.edu.sg

Abstract—With the advent of sophisticated artificial intelligence (AI) technologies, the proliferation of deepfakes and the spread of m/disinformation have emerged as formidable threats to the integrity of information ecosystems worldwide. This paper provides an overview of the current literature. Within the frontier AI’s crucial application in developing defense mechanisms for detecting deepfakes, we highlight the mechanisms through which generative AI based on large models (LM-based GenAI) craft seemingly convincing yet fabricated contents. We explore the multifaceted implications of LM-based GenAI on society, politics, and individual privacy violations, underscoring the urgent need for robust defense strategies. To address these challenges, in this study, we introduce an integrated framework that combines advanced detection algorithms, cross-platform collaboration, and policy-driven initiatives to mitigate the risks associated with AI-Generated Content (AIGC). By leveraging multi-modal analysis, digital watermarking, and machine learning-based authentication techniques, we propose a defense mechanism adaptable to AI capabilities of ever-evolving nature. Furthermore, the paper advocates for a global consensus on the ethical usage of GenAI and implementing cyber-wellness educational programs to enhance public awareness and resilience against m/disinformation. Our findings suggest that a proactive and collaborative approach involving technological innovation and regulatory oversight is essential for safeguarding netizens while interacting with cyberspace against the insidious effects of deepfakes and GenAI-enabled m/disinformation campaigns.

Index Terms—Deepfakes, disinformation, misinformation, large AI models, frontier AI, foundation models, AI-generated content (AIGC), generative AI.

I. INTRODUCTION

The frontier AI, characterized by its advanced capabilities and cutting-edge applications, significantly enhances the realism of deepfakes [1]. Concurrently, it is instrumental in devising innovative solutions to detect and counter m/disinformation. Frontier AI encompasses new, innovative AI technologies that could exhibit sufficiently dangerous capabilities such as generative AI, advanced machine learning algorithms, large models, etc. The implications of frontier AI technologies extend beyond technological advancements, necessitating a global consensus on ethical tools usage and the implementation of comprehensive cyber-wellness educational programs¹. Such measures are

critical in equipping society to navigate the complex dynamics of information dissemination and integrity in the era of frontier AI².

Over the last decade, the precipitous advancements in Generative Artificial Intelligence with large models (LM-based GenAI) have made revolutionary progress in crafting human-like multimedia content (e.g., text, image, video, or audio). Foundation models are a form of LM-based adaptable models that have become the backbone of significant technological progress, driving innovations from autonomous vehicles to personalized medicine [2]. However, considering the power of LM-based GenAI tools, they might bring unprecedented risks and unintended consequences to our society, for instance, by empowering malicious actors to apply for cyber-scramming or cyberbullying in the form of deepfake advertisements through social media platforms [3]. This paper delves into these phenomena by discussing both possible outcomes of LM-based GenAI models, their societal impacts, and the urgent need of today’s society for comprehensive defense mechanisms and sufficient cyber-wellness programs [4].

The rise of *Deepfake*, a portmanteau of “deep learning” and “fake” media, are digital fabrications in which realistic likenesses of things are synthetically generated or entirely altered to say or do something that never occurred [5]. Due to the public accessibility of sophisticated LM-based GenAI tools (e.g., ChatGPT and LivePerson), anyone can craft deepfake contents. As these capabilities become democratized, the potential for misuse scales exponentially. Mis/disinformation, closely related but not limited to deepfakes, encompasses all forms of false or misleading information deliberately spread to deceive netizens (active Internet users). This phenomenon is not new; however, the advent of LM-based GenAI models has supercharged its potential reach and believability. Multimedia contents (e.g., texts, images, audios, and videos) produced by the LM-based GenAI tools can now fabricate reality so that discovering truth from fiction becomes increasingly challenging (e.g., family voice cloning threats) [6].

¹<https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/>

²<https://www.channelnewsasia.com/singapore/ai-safety-summit-singapore-pm-lee-frontier-3892476>

The implications of the LM-based GenAI technologies are profound and multifaceted. Democracies worldwide grapple with the ramifications of AI generated content (AIGC) on electoral processes and public opinion [5]. Netizens face unprecedented threats to their privacy and security, as the deepfakes that might be created of their public data may act without their consent or even knowledge. Furthermore, the media landscape, the traditional outlets of factual information propagation, is undergoing a seismic shift as journalists and content creators confront the existential question of what is a piece of trustworthy information in the post-deepfake era [7].

II. BACKGROUND

This section gives the background for our discussion.

A. Historical Context of Information Manipulation

Historically, information manipulation was labor-intensive and required significant resources, restricting its practice to powerful entities such as state officials or large organizations [8]. The infamous propaganda of wartime misinformation campaigns, psychological operations, and political machinations are some testaments of how entities will affect public opinion or discredit opposition [9]. The advent of digital technology began a shift, enabling broader participation in information manipulation with the rise of Photoshop, video editing, social media platforms, and LM-based GenAI tools to disseminate such content widely and rapidly [10].

B. Frontier AI Amplifying and Combating Digital Deception

Frontier AI has reshaped the challenges in information manipulation. Its advances in neural networks and machine learning have heightened deepfakes realism, complicating the distinction between real and fake content³. Concurrently, frontier AI is crucial in developing tools to counter misinformation and disinformation, as highlighted in recent studies. This dual role underscores both its potential for generating and detecting digital falsehoods.

C. Evolution of LM-based GenAI Tools in Media Creation

The role of LM-based GenAI tools in media creation started benignly enough, with techniques designed to enhance image quality, recommend content, or power voice assistants. As machine learning models advanced, they transcended these supportive roles, becoming regular tools in content creation. Generative adversarial networks (GANs) [11], introduced in 2014, represented a significant leap forward, enabling the creation of photorealistic images indistinguishable from actual photographs by the unaided vision systems. The evolution of AI continues with LM-based GenAI that could synthesize human voices, compose music, and create realistic video footage [2].

D. AI-Generated Mis/Disinformation

Technically, if deepfakes are generated based on event-related concepts, they could be formed as mis/disinformation [12]. While text manipulation is less technologically complex than other media files, the implications are no less severe. Automated

ChatBots can disseminate false information by deploying LM-based GenAI tools that can craft fake news articles by claiming to be written by reputable sources. Malicious actors can deploy these ChatBots to spread using social media platforms, which can inadvertently prioritize and amplify misleading content.

E. Previous Efforts in Combating Digital Misinformation

In the literature, many researchers have taken promising steps to counter digital misinformation involving content moderation, community reporting, and algorithmic detection [2], [3]. However, these methods face challenges, such as the overwhelming content volume and evolving misinformation techniques. LM-based GenAI models play a significant role in spreading deepfakes that may cause mis/disinformation, necessitating a deeper understanding of effective defense strategies [13], [14].

III. THE RISE OF LARGE AI MODELS

The third decade of the 21st century is considered the landmark of a turning point in the capabilities of artificial intelligence, primarily through the advent of LM-based GenAI. AI foundation models and LM-based GenAI models (i.e., LLMs, LVMs, LAMs, or LMMs) have demonstrated unprecedented proficiency in understanding and generating human-like text, images, and sounds, leading to significant advancements in AIGC [15]. This section outlines the development of LM-based GenAI models, their capabilities, and their associated risks.

A. Overview of LM-based GenAI Models

LM-based GenAI models, such as OpenAI's GPT (Generative Pre-trained Transformer) series [16], Google's BERT (Bidirectional Encoder Representations from Transformers) [17], and others, represent the modern cutting edge technology, which craft contents automatically. These models include Large Language Models (LLMs), Large Vision Models (LVMs), Large Audio Models (LAMs), or Large Multimodal Models (LMMs) are characterized by their deep neural networks, which consist of millions or even billions of parameters that enable them to process and generate complex data patterns. The "large" in their name not only denotes their size in terms of parameters but also their vast training datasets and substantial computational power required for performing their operations.

B. Training and Functioning

The training process of the LM-based GenAI models involves feeding them enormous datasets, often sourced from the Internet, including books, articles, websites, and other publicly available media. This training allows the LM-based GenAI models to learn the nuances of human language, visual cues, and audio patterns [18]. They function by predicting the next word in a sentence, the next pixel in an image, or the following waveform in an audio file, learning from context and mimicking the style and texture of their training data [19].

C. Case Studies of Deepfakes and their Associated M/disinformation

Real-world instances of misusing LM-based GenAI technologies provide sobering case studies. Deepfake videos have been used to create fake celebrity advertisements, pornographic

³<https://www.gov.uk/government/publications/frontier-ai-taskforce-first-progress-report/frontier-ai-taskforce-first-progress-report>

videos, fabricated political speeches, and voice cloning of CEOs to commit fraud. AIGC has been employed to create fake news articles and social media posts that have gone viral, influencing public opinion and potentially affecting election outcomes [20].

D. Risks Associated with LM-based GenAI Capabilities

Evidently, the risks these models pose are beyond their capabilities as they provide opportunities for academic misconduct [21] or deepfake phishing [22] and many uncovered threats. The fact that LM-based GenAI tools can be deployed in deceptive scenarios as convincing mis/disinformation provides opportunities for virtually anyone with the requisite technical know-how to launch sophisticated misleading campaigns. The potential for these technologies to be used for blackmail, electoral interference, and social unrest is a pressing concern [23]. Moreover, the speed at which AIGC can be produced outstrips the ability of current detection and moderation systems to keep up, creating a game of digital cat-and-mouse where the mouse is increasingly agile.

IV. SOCIETAL IMPLICATIONS

The societal implications of deepfakes and mis/disinformation generated by LM-based GenAI are bringing unprecedented impacts, touching upon every facet of modern life—from politics and security to individual rights and societal trust [24]. In the following, we provide an overview of the far-reaching consequences of these phenomena and underscore the critical need for a robust societal response.

A. Effects on Democracy and Public Opinion

In democratic societies, the integrity of public discourse is foundational. Deepfakes can be deployed as LM-based GenAI-generated mis/disinformation that threaten the integrity of news propagation, as they could be exploited for fabricating scandals, falsifying records of public statements, and manipulating electoral processes. When voters cannot distinguish between real and falsified representations of candidates or policies, the very fabric of democratic decision-making is undermined [25]. The dissemination of spurious information can sway elections, fuel political polarization, and erode the public’s trust in democratic governments.

B. Impact on Privacy and Personal Security

The ability to create convincing fake images and videos of individuals without getting their consent has raised alarm bells regarding privacy and personal security. Deepfakes can be weaponized to discredit individuals, exploit them for blackmail, or invade their privacy in egregious ways, as seen in the creation of non-consensual deepfake pornography [26]. The impacts of such artifacts are deceptive effects on free expression and the pervasive sense of vulnerability as individuals grapple with the potential for their likeness to be used in harmful ways.

C. Consequences for Media and Journalism

Technically, Journalism’s role as the fourth estate is predicated on the ability to provide accurate, reliable information. Deepfakes and AIGC pose existential challenges to this role. Journalists are forced to contend with the additional burden of

verifying content authenticity while the public grows increasingly skeptical of media reports [27]. This skepticism can lead to a ‘cry wolf’ scenario, where even legitimate news contents are doubted, contributing to a disconcerting post-truth era where facts are fungible, and the truth is subjective.

D. Erosion of Public Trust

The cumulative effects of unchecked deepfakes and misinformation are the erosion of public trust [28]. When netizens cannot trust their eyes or ears, they can become cynical and disengaged. This disengagement poses risks not just to political processes but to the social fabric that binds communities together. Without trust, conspiracy theories flourish, scientific consensus is questioned, and social polarization deepens.

E. Legal and Ethical Dilemmas

The rise of AIGC has also precipitated legal and ethical dilemmas [29]. Current laws are ill-equipped to handle the nuances of deepfakes, often lagging behind technological advancements. Ethically, the implications are just as complex as creating and distributing deepfakes of people without their consent, violating their rights.

V. TECHNICAL DEFENSE MECHANISMS

This section discusses the technological, strategic, and policy-oriented defense approaches that can mitigate the risks associated with AIGC. Since the realistic construction of deepfakes and dissemination of mis/disinformation have become more sophisticated with the advancement of LM-based GenAI tools, developing robust technical defense mechanisms is a complex agenda. Below, we outline current and emerging technologies aimed at detecting and countering AIGC and the challenges inherent in their deployment.

A. Detection Algorithms

Detection is the first line of defense against AI-generated false content. Algorithms designed to identify deepfakes typically analyze various data points that may indicate manipulation, such as inconsistencies in lighting, unnatural blinking patterns, or irregularities in skin texture. Advances in machine learning have led to the development of models that can scrutinize video frames for signs of alteration at a pixel level, often with the aid of deep learning techniques similar to those used to create deepfakes [3]. Audio deepfake detection similarly analyzes vocal patterns, looking for subtle signs of manipulation that may not be apparent to the human ear. These include irregularities in speech patterns, breathing sounds, and background noises [3]. The challenge lies in the fact that as detection algorithms become more sophisticated, so too do the methods for creating deepfakes, leading to an ongoing arms race between creators and detectors.

B. AI-Driven Authentication Methods

In addition to detection, authentication methods aim to verify the origin and integrity of content. Digital watermarking, for instance, involves embedding a hidden and unique pattern or code within the content at the time of creation, which can later be used to confirm its authenticity [30]. Blockchain technology

offers another layer of security by providing a decentralized and immutable ledger of content creation and distribution, making unauthorized alterations easily traceable. Another approach is the use of biometric authentication, which employs unique biological characteristics such as facial recognition patterns, voiceprints, or even typing rhythms to confirm the identity of individuals in digital media [31]. These methods, however, must balance the need for security with concerns about privacy and the potential for misuse.

C. Machine Learning-Based Authentication Techniques

Machine learning is not only used to create deepfakes but can also be harnessed to combat them. Models can be trained to recognize the digital ‘fingerprints’ left by the AI models that generate deepfakes. These fingerprints are often subtle flaws or patterns in the generated content that are consistent with the training data or generation method used [32]. By analyzing these fingerprints, machine learning algorithms can identify whether content has been artificially generated or altered.

D. Limitations and Challenges of Current Technologies

While these technologies show promise, they are not without limitations. Deepfake creation techniques are evolving rapidly, and detection methods must continually adapt to keep pace [33]. Moreover, the computational resources required to analyze large volumes of content in real time are substantial, and false positives remain a concern. Another challenge is the ease of access to deepfake generation tools, which can be used by individuals with minimal technical expertise, further complicating detection efforts [34]. Additionally, the adaptability of AI means that as soon as a detection method becomes effective, new techniques are developed to circumvent it. This cat-and-mouse dynamic requires a proactive and dynamic approach to defense mechanism development.

E. The Need for Open Collaboration

Given the scale and complexity of the challenge, open collaboration between academia, industry, and government is necessary. Sharing data, research findings, and strategies can accelerate the development of effective defense mechanisms [35]. Transparency in the functioning of detection and authentication technologies is also crucial to build trust and ensure these tools are used responsibly.

VI. CROSS-PLATFORM STRATEGIES

The digital ecosystem’s interconnected nature necessitates cross-platform strategies to combat the spread of deepfakes and mis/disinformation effectively. This section outlines a collaborative approach that spans various stakeholders, including social media companies, technology firms, content creators, and end-users.

A. The Role of Social Media and Technology Companies

Social media platforms are the primary battlegrounds for the spread of deepfakes and mis/disinformation due to their vast reach and the speed at which content can go viral. These companies have a responsibility to actively monitor and mitigate the spread of fake content. Strategies include [36]:

- Content Moderation Enhancements: Using a combination of AI-driven and human moderation to detect and flag deepfakes.
- Partnerships with Fact-Checkers: Collaborating with independent fact-checking organizations to verify content.
- User Reporting Mechanisms: Empowering users to report suspicious content, which can then be reviewed by specialized teams.
- Transparency Reports: Publishing regular reports on the number of deepfakes detected and the actions taken.
- User Education: Providing educational resources to help users spot and understand the nature of deepfakes.

B. Collaborative Filtering and Fact-Checking Initiatives

Collaborative filtering involves leveraging the collective effort of platform users to identify and filter out disinformation [37]. This can be facilitated through:

- Community-Driven Moderation: Enabling community moderators to review and moderate content within their domains of expertise.
- Crowdsourced Verification: Utilizing crowdsourcing to gather user input on the authenticity of content.
- Real-Time Fact-Checking: Implementing systems that provide live fact-checking during events, speeches, and debates.

C. User-Centric Approaches

Putting users at the center of the defense strategy involves education and empowerment [38]. This includes:

- Digital Literacy Programs: Educating the public on digital media, the existence of deepfakes, and the importance of critical thinking online.
- Critical Media Literacy: Encouraging users to question the source and intent behind the content they consume.
- Promotion of Verified Content: Boosting the visibility of content from verified and reputable sources.

D. Community Guidelines and Enforcement

Platforms must establish clear community guidelines that define acceptable use and the consequences of spreading deepfakes and mis/disinformation [39]. Enforcement actions may include:

- Content Removal: Removing or demoting content that violates platform policies.
- Account Suspension: Temporarily or permanently suspending accounts that repeatedly disseminate fake content.
- User Feedback: Informing users when they have interacted with or shared false content.

E. Developing Standardized Protocols

To streamline cross-platform efforts, there is a need for standardized protocols for content verification, data sharing, and incident response. This could involve [40]:

- Interoperable Verification Tags: Creating tags that indicate content has been verified, which can be recognized across different platforms.
- Data Sharing Agreements: Establishing agreements to share data on deepfakes and misinformation trends and techniques.

- Joint Response Frameworks: Developing coordinated response plans for widespread disinformation campaigns.

VII. ETHICAL CONSIDERATIONS

The ethical implications of deepfakes and misinformation are as vast and complex as their technical and social counterparts [41]. This section explores the moral landscape that AIGC presents, the responsibilities of creators and disseminators, and the overarching need for ethical guidelines to shape the evolution of AI technologies.

A. Ethical AI Development and Use

The development of AI technologies is not value-neutral; it reflects the biases, priorities, and ethical orientations of its creators. Therefore, the following needs to be addressed.

- Bias and Fairness: There is a need for ethical AI development that actively seeks to minimize biases in training data and algorithms, ensuring fairness and non-discrimination [42].
- Transparency: AI systems should be developed with transparency in mind, allowing for traceability and explainability in the AI's decision-making processes [43].
- Accountability: Developers and users of AI must be accountable for the outcomes of their technologies, particularly when they impact public opinion or infringe on personal rights [44].

B. The Balance between Innovation and Regulation

There is a delicate balance to be maintained between encouraging innovation in AI and implementing regulations that protect against its misuse:

- Innovation-Friendly Policies: Policies should aim to foster innovation and the beneficial applications of AI while guarding against risks.
- Proactive Ethical Design: AI should be designed proactively with ethical considerations in mind, rather than retroactively applying ethical standards to existing technologies.

C. Future Outlook and Philosophical Implications

AI's capabilities force us to confront deep philosophical questions about the nature of truth, reality, and human experience:

- Ontological Questions: As AI blurs the lines between reality and simulation, we must address the ontological status of experiences and entities created by AI.
- Epistemological Considerations: The proliferation of deepfakes calls into question the basis of knowledge and the conditions under which we can claim to know something as true or false.
- Human Agency and Autonomy: There is a need to consider how AI impacts human agency and autonomy, particularly when individuals are subject to AI-generated representations without their consent.

D. The Ethical Use of Deepfakes

While deepfakes are often discussed in negative terms, they also have potentially positive applications:

- Artistic and Educational Uses: Deepfakes can be used for legitimate artistic expression or educational purposes, such as recreating historical speeches [5].
- Medical and Therapeutic Applications: There are possibilities for using deepfake technology in medical simulations or therapeutic settings [45].

VIII. PROPOSED INTEGRATED DEFENSE FRAMEWORK

The multifaceted nature of the threats posed by deepfakes and mis/disinformation necessitates a comprehensive response [46]. This section proposes an integrated defense framework that synthesizes technological, strategic, policy-oriented, and educational responses to these threats.

A. Design of the Integrated Defense Framework

The proposed framework is designed with four key pillars:

- Technological Solutions: Incorporating advanced detection algorithms, AI-driven authentication methods, and machine learning-based authentication techniques.
- Strategic Initiatives: Implementing cross-platform strategies, including content moderation enhancements and collaborative filtering.
- Policy and Regulation: Developing new legislation and ethical guidelines that clearly define and impose penalties for the creation and distribution of deepfakes.
- Education and Public Awareness: Launching comprehensive educational programs and public awareness campaigns to improve media literacy and critical thinking.

B. Implementation of the Framework

For effective implementation, the framework requires:

- Multi-Stakeholder Collaboration: Coordination among governments, tech companies, academia, and civil society to ensure a united front against deepfakes.
- Resource Allocation: Commitment of financial, human, and technological resources to support the framework's initiatives.
- Adaptive Strategies: Continuous adaptation of strategies to address the evolving nature of deepfake and misinformation tactics [47].

C. Case Study: Applying the Framework in a Simulated Environment

To validate the framework, a simulated environment that replicates the complex ecosystem of media platforms and AIGC can be created [48]. Here, the framework's components would be tested against various attack scenarios to assess their effectiveness and identify areas for improvement.

D. Analysis of Framework Effectiveness

Evaluating the effectiveness of the defense framework involves:

- **Monitoring and Evaluation:** Regular assessment of each pillar's performance in detecting and countering deepfakes.
- **Feedback Mechanisms:** Systems for collecting feedback from stakeholders to inform the iterative improvement of the framework.
- **Benchmarking:** Setting benchmarks for success and conducting comparative analysis with other defense strategies.

E. Potential Unforeseen Consequences and Mitigation Strategies

While the framework aims to be comprehensive, there may be unforeseen consequences, such as over-censorship or the stifling of innovation. Mitigation strategies include:

- **Ethical Oversight:** Establishing ethical oversight committees to review the impact of defense measures.
- **Balanced Approach:** Ensuring a balanced approach that respects freedom of expression while protecting against misinformation.
- **Rapid Response Protocols:** Developing protocols for rapidly addressing negative consequences as they arise.

IX. DISCUSSION

The emergence of deepfakes and the proliferation of mis/disinformation through the use of advanced AI models pose a significant threat to the integrity of information, necessitating a multi-pronged approach to mitigation [49]. This discussion evaluates the proposed solutions, explores potential unintended consequences, and highlights ongoing challenges and areas for future research.

A. Analysis of the Proposed Solutions' Effectiveness

The proposed integrated framework's effectiveness hinges on the synergy between its components:

- **Technological Efficacy:** The rapid detection of deepfakes is crucial. However, as the technology to create deepfakes becomes more sophisticated, detection methods may need to become more specialized, potentially leading to an arms race between creation and detection capabilities.
- **Strategic Resilience:** Cross-platform strategies emphasize the need for a coordinated response to misinformation. The scalability of such initiatives is vital, as is the ability to adapt quickly to new forms of disinformation.
- **Policy Impact:** The effectiveness of policy measures will largely depend on their enforcement and the international community's willingness to adopt and implement harmonized standards.
- **Educational Outcomes:** Long-term, the success of educational programs in enhancing the public's ability to discern true from false information may be one of the most sustainable defenses against misinformation.

B. Open Challenges and Areas for Future Research

Several challenges remain open, requiring ongoing attention:

- **Technological Advancement:** Keeping defensive measures up-to-date with the latest advancements in AI and deepfake technologies.
- **Global Cooperation:** Achieving a consensus on international standards and cooperation in the face of geopolitical tensions and differing national interests.
- **Public Engagement:** Ensuring continued public engagement and understanding in the face of "fatigue" around the topic of misinformation. Future research areas are plentiful, including:
- **Behavioral Insights:** Gaining a deeper understanding of why people create and spread misinformation, and how they are influenced by it.
- **Economic Models:** Developing economic models to understand the incentives behind the spread of deepfakes and misinformation [50].
- **Technological Innovations:** Exploring new technological innovations that can preemptively address the creation of deepfakes.

X. CONCLUSION

The paper emphasizes the critical role of frontier AI in countering the profound threat of deepfakes and generative AI to global information ecosystems. It underscores the need for a comprehensive, multi-faceted defense strategy that evolves in tandem with frontier AI advancements. The paper highlights the importance of developing sophisticated technological solutions, adaptable international policies, and enhancing public education in media literacy to effectively combat these threats. Advocating for a collaborative approach, it integrates the advancements in frontier AI with regulatory strategies and media literacy education, framing the battle against deepfakes as not only a technical challenge but a broader societal issue.

ACKNOWLEDGEMENT

This research is partly supported by the Singapore Ministry of Education Academic Research Fund under Grant Tier 1 RG90/22, Grant Tier 1 RG97/20, Grant Tier 1 RG24/20 and Grant Tier 2 MOE2019-T2-1-176; and partly by the Nanyang Technological University (NTU)-Wallenberg AI, Autonomous Systems and Software Program (WASP) Joint Project.

REFERENCES

- [1] M. Anderljung, J. Barnhart, J. Leung, A. Korinek, C. O'Keefe, J. Whittlestone, S. Avin, M. Brundage, J. Bullock, D. Cass-Beggs *et al.*, "Frontier ai regulation: Managing emerging risks to public safety," *arXiv preprint arXiv:2307.03718*, 2023.
- [2] Y. Cao, S. Li, Y. Liu, Z. Yan, Y. Dai, P. S. Yu, and L. Sun, "A comprehensive survey of AI-generated content (AIGC): A history of generative AI from GAN to ChatGPT," *arXiv preprint arXiv:2303.04226*, 2023.
- [3] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 1, pp. 1–41, 2021.
- [4] W. Shin and M. O. Lwin, "Parental mediation of children's digital media use in high digital penetration countries: perspectives from singapore and australia," *Asian Journal of Communication*, vol. 32, no. 4, pp. 309–326, 2022.

- [5] V. Danry, J. Leong, P. Pataranutaporn, P. Tandon, Y. Liu, R. Shilkrot, P. Punpongsonan, T. Weissman, P. Maes, and M. Sra, "AI-generated characters: putting deepfakes to good use," in *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 2022, pp. 1–5.
- [6] N. Amezaga and J. Hajek, "Availability of voice deepfake technology and its impact for good and evil," in *Proceedings of the 23rd Annual Conference on Information Technology Education*, 2022, pp. 23–28.
- [7] J. Fletcher, "Deepfakes, artificial intelligence, and some kind of dystopia: The new faces of online post-fact performance," *Theatre Journal*, vol. 70, no. 4, pp. 455–471, 2018.
- [8] R. W. Zmud, "Opportunities for strategic information manipulation through new information technology," *Organizations and Communication Technology*, pp. 95–116, 1990.
- [9] D. Silverman, K. Kaltenthaler, and M. Dagher, "Seeing is disbelieving: the depths and limits of factual misinformation in war," *International Studies Quarterly*, vol. 65, no. 3, pp. 798–810, 2021.
- [10] M. T. Ahvanooy, Q. Li, X. Zhu, M. Alazab, and J. Zhang, "ANiTW: A novel intelligent text watermarking technique for forensic identification of spurious information on social media," *Computers & Security*, vol. 90, p. 101702, 2020.
- [11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, 2014.
- [12] J. Zhou, Y. Zhang, Q. Luo, A. G. Parker, and M. De Choudhury, "Synthetic lies: Understanding AI-generated misinformation and evaluating algorithmic and human solutions," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–20.
- [13] B. He, Y. Hu, Y. Lee, S. Oh, G. Verma, and S. Kumar, "A survey on the role of crowds in combating online misinformation: Annotators, evaluators, and creators," *arXiv*, 2023, accessed November 21, 2023.
- [14] S. Siwakoti, J. N. Shapiro, and N. Evans, "Less reliable media drive interest in anti-vaccine information," *Harvard Kennedy School Misinformation Review*, 2023.
- [15] M. U. Hadi, R. Qureshi, A. Shah, M. Irfan, A. Zafar, M. B. Shaikh, N. Akhtar, J. Wu, S. Mirjalili *et al.*, "Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects," 2023.
- [16] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.
- [17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [18] S. Biderman, H. Schoelkopf, Q. G. Anthony, H. Bradley, K. O'Brien, E. Hallahan, M. A. Khan, S. Purohit, U. S. Prashanth, E. Raff *et al.*, "Pythia: A suite for analyzing large language models across training and scaling," in *International Conference on Machine Learning*. PMLR, 2023, pp. 2397–2430.
- [19] G. Xiao, J. Lin, M. Seznec, H. Wu, J. Demouth, and S. Han, "Smoothquant: Accurate and efficient post-training quantization for large language models," in *International Conference on Machine Learning*. PMLR, 2023, pp. 38 087–38 099.
- [20] D. Xu, S. Fan, and M. Kankanhalli, "Combating misinformation in the era of generative AI models," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 9291–9298.
- [21] S. A. Bin-Nashwan, M. Sadallah, and M. Bouteraa, "Use of chatgpt in academia: Academic integrity hangs in the balance," *Technology in Society*, vol. 75, p. 102370, 2023.
- [22] Y. Mirsky, A. Demontis, J. Kotak, R. Shankar, D. Gelei, L. Yang, X. Zhang, M. Pintor, W. Lee, Y. Elovici *et al.*, "The threat of offensive ai to organizations," *Computers & Security*, vol. 124, p. 103006, 2023.
- [23] M. Mustak, J. Salminen, M. Mäntymäki, A. Rahman, and Y. K. Dwivedi, "Deepfakes: Deceptions, mitigations, and opportunities," *Journal of Business Research*, vol. 154, p. 113368, 2023.
- [24] S. Gregory, "Fortify the truth: How to defend human rights in an age of deepfakes and generative ai," p. huad035, 2023.
- [25] K. J. Schiff, D. S. Schiff, and N. Bueno, "The liar's dividend: The impact of deepfakes and fake news on trust in political discourse," 2023.
- [26] U. A. Ciftci, G. Yuksek, and I. Demir, "My face my choice: Privacy enhancing deepfakes for social media anonymization," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 1369–1379.
- [27] K. Wahl-Jorgensen and M. Carlson, "Conjecturing fearful futures: journalistic discourses on deepfakes," *Journalism Practice*, vol. 15, no. 6, pp. 803–820, 2021.
- [28] M. Pawelec, "Deepfakes and democracy (theory): how synthetic audiovisual media for disinformation and hate speech threaten core democratic functions," *Digital Society*, vol. 1, no. 2, p. 19, 2022.
- [29] C. Öhman, "Introducing the pervert's dilemma: a contribution to the critique of deepfake pornography," *Ethics and Information Technology*, vol. 22, no. 2, pp. 133–140, 2020.
- [30] Y. Wang, "Synthetic realities in the digital age: Navigating the opportunities and challenges of ai-generated content," 2023.
- [31] C. Campbell, K. Plangger, S. Sands, and J. Kietzmann, "Preparing for an era of deepfakes and ai-generated ads: A framework for understanding responses to manipulated advertising," *Journal of Advertising*, vol. 51, no. 1, pp. 22–38, 2022.
- [32] A. Mitra, S. P. Mohanty, P. Corcoran, and E. Kougiyanos, "A machine learning based approach for deepfake detection in social media through key video frame extraction," *SN Computer Science*, vol. 2, pp. 1–18, 2021.
- [33] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, "Multi-attentional deepfake detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2185–2194.
- [34] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, "Deepfake generation and detection: State-of-the-art, open challenges, countermeasures, and way forward," *Applied Intelligence*, vol. 53, no. 4, pp. 3974–4026, 2023.
- [35] D. Krishna, "Deepfakes, online platforms, and a novel proposal for transparency, collaboration, and education," *Rich. J.L. & Tech.*, vol. 27, p. 1, 2020.
- [36] J. T. Hancock and J. N. Bailenson, "The social impact of deepfakes," pp. 149–152, 2021.
- [37] A. Ünver, "Emerging technologies and automated fact-checking: Tools, techniques and algorithms," *Techniques and Algorithms (August 29, 2023)*, 2023.
- [38] P. Gupta, K. Chugh, A. Dhall, and R. Subramanian, "The eyes know it: Fake-an eye-tracking database to understand deepfake perception," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, 2020, pp. 519–527.
- [39] K. Kikerpill, A. Siibak, and S. Valli, "Dealing with deepfakes: Reddit, online content moderation, and situational crime prevention," in *Theorizing Criminality and Policing in the Digital Media Age*. Emerald Publishing Limited, 2021, vol. 20, pp. 25–45.
- [40] K. Shiohara and T. Yamasaki, "Detecting deepfakes with self-blended images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 720–18 729.
- [41] N. Diakopoulos and D. Johnson, "Anticipating and addressing the ethical implications of deepfakes in the context of elections," *New Media & Society*, vol. 23, no. 7, pp. 2072–2098, 2021.
- [42] B. Giovanola and S. Tiribelli, "Beyond bias and discrimination: redefining the AI ethics principle of fairness in healthcare machine-learning algorithms," *AI & Society*, vol. 38, no. 2, pp. 549–563, 2023.
- [43] U. Ehsan, Q. V. Liao, M. Muller, M. O. Riedl, and J. D. Weisz, "Expanding explainability: Towards social transparency in AI systems," in *ACM CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–19.
- [44] A. Henriksen, S. Enni, and A. Bechmann, "Situating accountability: Ethical principles, certification standards, and explanation methods in applied AI," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021, pp. 574–585.
- [45] H.-C. Yang, A. R. Rahmanti, C.-W. Huang, and Y.-C. J. Li, "How can research on artificial empathy be enhanced by applying deepfakes?" *Journal of Medical Internet Research*, vol. 24, no. 3, p. e29506, 2022.
- [46] D. Kelly and J. Burkell, "It's not (all) about the information: The role of cognition in creating and sustaining false beliefs," *Cambridge Studies on Governing Knowledge Commons*, 2024.
- [47] P. T. Jaeger and N. G. Taylor, "Arsenals of lifelong information literacy: Educating users to navigate political and current events information in world of ever-evolving misinformation," *The Library Quarterly*, vol. 91, no. 1, pp. 19–31, 2021.
- [48] Y. Liu, H. Du, D. Niyato, J. Kang, Z. Xiong, C. Miao, and A. Jamalipour, "Blockchain-empowered lifecycle management for AI-generated content (AIGC) products in edge networks," *arXiv preprint arXiv:2303.02836*, 2023.
- [49] K. Langmia, *Black Communication in the Age of Disinformation: Deep-Fakes and Synthetic Media*. Springer Nature, 2023.
- [50] N. Kshetri, "The economics of deepfakes," *Computer*, vol. 56, no. 8, pp. 89–94, 2023.

This figure "fig1.png" is available in "png" format from:

<http://arxiv.org/ps/2311.17394v1>