I've Got Your Daughter: Generative AI Boosts Deepfake-Enhanced Cybercrime

ZIYI ZHAO, Fox School of Business, Temple University, USA

The phenomenon of deepfake, leveraging generative AI to create hyper-realistic videos and audio recordings, presents a formidable challenge in the digital age, blurring the lines between reality and fabrication with implications for privacy, security, and trust. This study addresses the critical research question of how deepfake technology affects the detection of cybercrime, particularly through the impersonation of individuals. Grounded in Interpersonal Deception Theory (IDT), the research modifies traditional constructs to encompass the unique characteristics of deepfakes. The empirical investigation faces hurdles in acquiring authentic deepfake data and ethical concerns, addressed through a mixed-method approach including sentiment analysis of online videos, controlled experiments, and semi-structured interviews. Preliminary findings indicate varied public perceptions of deepfakes, influenced significantly by the content's context. This research contributes to the HCI field by elucidating the dynamic interaction between deepfake technology and human cognition, offering insights for developing more effective cybercrime detection and prevention strategies.

CCS Concepts: • **Information systems** • **Social and professional topics** • **General and reference**

KEYWORDS: Generative AI, deepfake, cybercrime, cyber security

ACM Reference format:

Ziyi Zhao. 2024. I've Got Your Daughter: Generative AI Boosts Deepfake-Enhanced Cybercrime. *Generative AI in User-generated Content Workshop. In Proceedings of the ACM on Human-Computer Interaction.*

EXTENDED ABSTRACT

"Mom, these bad men have me: She believes scammers cloned her daughter's voice in a fake kidnapping."

– a headline from CNN News [1]

Imagine a scenario where you receive an urgent phone call from your child. The voice is one you would recognize at a glance. It is a convincing description of a situation full of genuine fear and panic. Would you be able to think calmly, or would your instincts prompt an immediate response? The sad reality is that such scenarios, which prey on our deepest fears and emotions, have moved from nightmares to the real world. In recent news, AI has been used by scammers in the U.S. to imitate the voice of a teenage girl and trick her mother into paying a million-dollar ransom.[1] In another country, a scammer used AI to convincingly pose as a friend of the victim's via a face-swapped video call, resulting in the transfer of approximately $610,000.[2]

These cases illustrate the misuse of AI technology known as "deepfake", which has become a growing concern for law enforcement (Caldwell et al., 2020). Deepfake utilizes AI to manipulate digital content, such as videos and images, to create highly realistic representations of people performing actions they never did (Westerlund, 2019). Several

[1] https://www.cnn.com/2023/04/29/us/ai-scam-calls-kidnapping-cec/index.html
[2] https://www.reuters.com/technology/deepfake-scam-china-fans-worries-over-ai-driven-fraud-2023-05-22/

years ago, the creation of deepfakes was limited to a select few due to their high complexity, and the results were not entirely convincing (Yu et al., 2022). Nowadays, however, rapid advances in generative AI (GenAI) have made deepfake remarkably realistic.

A GenAI-supported deepfake can convincingly and inexpensively replicate an individual's appearance, voice, and mannerisms, which makes it very difficult to detect.[3] Remarkably, GenAI can interact with users using the content and style it generates, even if that information is not available in the original data used for training.[4] This development positions deepfake as a powerful tool in the hands of cybercriminals. As risks increase, it is crucial that this issue be thoroughly researched and effective countermeasures be implemented.

Traditionally, the focus of cybercrime research has been on text-based crimes such as phishing. Phishing is using deceptive emails to trick individuals into revealing sensitive information (Wright et al., 2023). However, the rise of deepfakes has been a game-changer in this landscape. Audio and video are considered to be richer forms of media than text because they provide immediate feedback and can convey complex messages (Daft and Lengel, 1986). While individuals may approach text-based scams cautiously, they are more susceptible to communication that mimics a familiar voice or face (Hancock et al., 2004). The seriousness of this threat is underscored by the fact that more than half of the 200,000 monthly "unwanted call" complaints[5] in the U.S. are robocalls, the majority of which are impersonations. This evolving context calls for a significant shift in research focuses beyond traditional text-based cybercrime to address the emerging threat of deepfake-enhanced audio and video cybercrime.

The emergence and ongoing development of deepfake underscores the urgent need for a broader discourse on digital responsibility that incorporates societal and ethical considerations. The U.S. National Strategic Plan for Artificial Intelligence Research and Development advocates for a sociotechnical approach to real-world applications considering the political and organizational responsibilities governing human-AI interactions (National Science and Technology Council, 2023). As a result, it is incumbent upon IS researchers to adopt the sociotechnical perspective and address the dynamic interaction between deepfake and humans. This approach should go beyond the static, one-sided concept of deception and allow for a comprehensive understanding of this complex issue. This study, therefore, investigates the following research question: *How does the deepfake, particularly the impersonation of intimates, affect cybercrime detection?*

[3] https://mashable.com/article/moments-ai-fooled-internet-deepfakes-misinformation
[4] https://www.techtarget.com/searchenterpriseai/definition/generative-AI
[5] https://consumer.ftc.gov/unwanted-calls-emails-and-texts/unwanted-calls

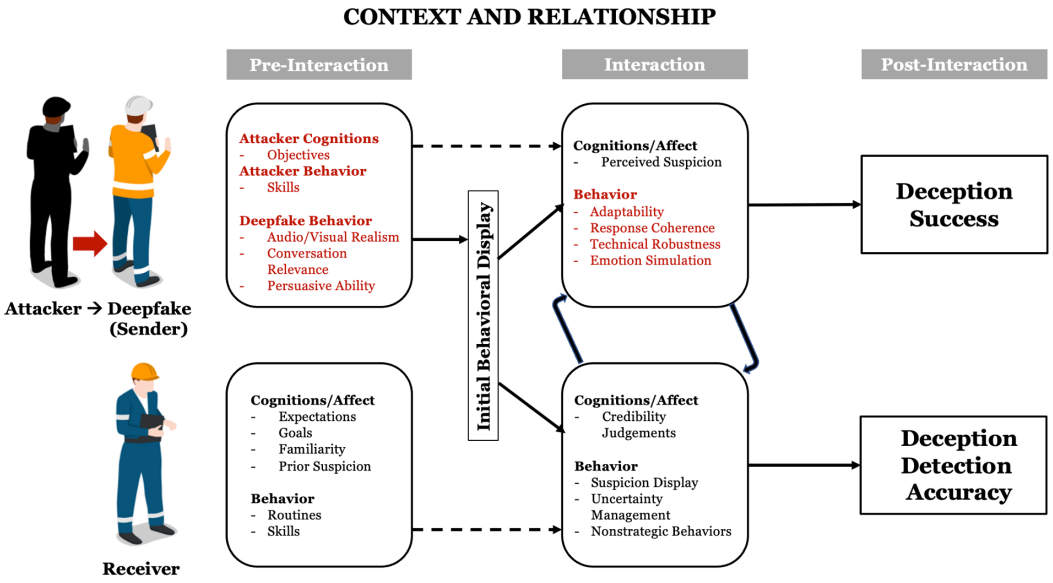

Fig. 1. Adapted Interpersonal Deception Theory in the Deepfake Context

Theoretically, we contextualize Interpersonal Deception Theory (IDT, Buller and Burgoon, 1996) by modifying the key constructs of the theory to account for the unique nature of deepfake, therefore proposing an *adopted IDT framework* (shown in Figure 1, with red fonts the new constructs proposed by this study). This involves considering the "sender" as both the human attacker and the deepfake, thus extending the human-centered theory to a technology-driven context. Furthermore, we propose a novel feedback loop within our conceptual model that reflects the dynamic, interactive nature of deepfake deception (shown in Figure 2). This loop encapsulates the continuous interplay between the deepfake and the deception receiver and demonstrates the role of GenAI in enhancing the credibility of the deepfake.

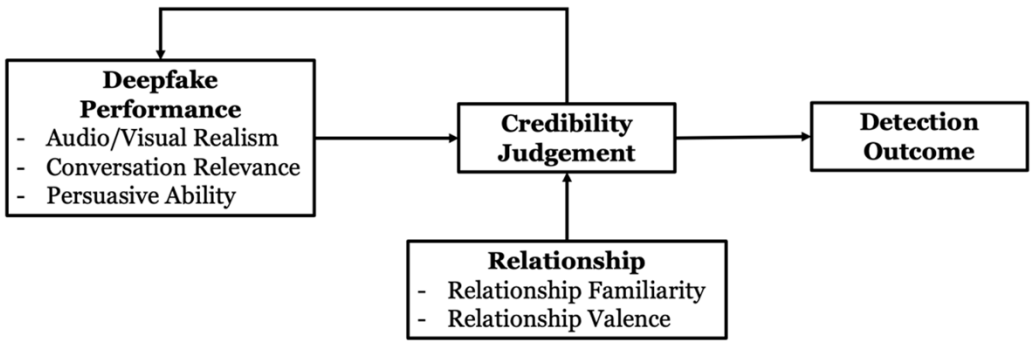

Fig. 2. Conceptual Model

The empirical challenges in this research are twofold: 1) obtaining public records of actual deepfake voice or video calls and data of attackers and victims is currently not feasible, and 2) the potential psychological and emotional distress that may arise if we design and conduct experiments with a deepfake embedded. To address these challenges, we have developed a three-step method. First, we scrape publicly available deepfake-related videos from YouTube and conduct sentiment analysis of video comments (Lu et al., 2022). This assists in understanding the deepfake performance construct in our research model. Second, we conduct an online experiment by extending the methods used in text-based phishing research (e.g., Chen et al., 2020; Wright et al., 2014) to a video-based phishing scenario, which allows us to investigate the detection outcome and potential mechanisms. However, these methods are insufficient for studying the feedback loop through which deepfake learns from its victims in real time, impacting judgments and detections iteratively. To gain a better understanding of this dynamic from multiple stakeholders, semi-interviews with approximately victims of deepfake cybercrime, law enforcement officials, and cybercrime and cybersecurity experts (Jha et al. 2016) would be a useful supplement.

We conducted a sentiment analysis by scraping the 82 most popular YouTube videos tagged with "deepfake," utilizing the YouTube Data API. Initial analysis based on a sentiment analysis method revealed significant variations in sentiments across different video topics: entertainment-focused deepfake videos often elicited positive sentiments about realism and persuasiveness, whereas videos on politics or technology garnered mixed reactions, from admiration of technological advancement to concerns over ethics and crimes. However, this approach struggled to capture the nuanced and mixed sentiments within individual comments, prompting us to explore a deep learning-based Aspect-Based Sentiment Analysis (ABSA) approach for our next steps (Chauhan et al., 2023; Wang et al., 2021). This advanced method overcomes traditional sentiment analysis's limitations, such as lower precision and the need for updates to the emotional dictionary, by extracting sentiments related to specific aspects like realism, relevance, and persuasiveness.

Our study contributes to the literature by examining the interactive dynamics surrounding deepfake technologies, a facet that has been underexplored in the current literature which primarily emphasizes the technical aspects of deepfake creation and detection (e.g., Heidari et al., 2023; Ju et al., 2024). By analyzing how individuals perceive and engage with deepfake content, our research offers valuable insights for developing targeted educational initiatives that inform potential victims about deepfake threats. Furthermore, our findings contribute to the formulation of effective regulatory frameworks and legal guidelines, thereby enhancing cybercrime prevention measures.

REFERENCES

Buller, D. B., & Burgoon, J. K. (1996). Interpersonal deception theory. *Communication theory*, *6*(3), 203-242.

Caldwell, M., Andrews, J. T., Tanay, T., & Griffin, L. D. (2020). AI-enabled future crime. *Crime Science*, *9*(1), 1-13.

Chauhan, G. S., Nahta, R., Meena, Y. K., & Gopalani, D. (2023). Aspect based sentiment analysis using deep learning approaches: A survey. *Computer Science Review*, *49*, 100576.

Chen, R., Gaia, J., & Rao, H. R. (2020). An examination of the effect of recent phishing encounters on phishing susceptibility. *Decision Support Systems, 133*, 1-14. http://dx.doi.org/10.1016/ j.dss.2020.113287

Daft, R. L., & Lengel, R. H. (1986). Organizational information requirements, media richness and structural design. *Management science*, *32*(5), 554-571.

Hancock, J. T., Thom-Santelli, J., & Ritchie, T. (2004). Deception and design: The impact of communication technology on lying behavior. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 129-134).

Heidari, A., Jafari Navimipour, N., Dag, H., & Unal, M. (2023). Deepfake detection using deep learning methods: A systematic and comprehensive review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, e1520.

Jha, S. K., Pinsonneault, A., & Dubé, L. (2016). The evolution of an ict platform-enabled ecosystem for poverty alleviation. *MIS Quarterly*, *40*(2), 431-446.

Ju, Y., Hu, S., Jia, S., Chen, G. H., & Lyu, S. (2024). Improving fairness in deepfake detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 4655-4665).

Lu, Y., Wu, J., Tan, Y., & Chen, J. (2022). Microblogging Replies and Opinion Polarization: A Natural Experiment. *MIS Quarterly*, *46*(4).

National Science and Technology Council (2023). *National Artificial Intelligence Research and Development Strategic Plan: 2023 Update*. Office of Science and Technology Policy. https://www.whitehouse.gov/wp-content/uploads/2023/05/National-Artificial-Intelligence-Research-and-Development-Strategic-Plan-2023-Update.pdf

Wang, J., Xu, B., & Zu, Y. (2021, July). Deep learning for aspect-based sentiment analysis. In *2021 International Conference on Machine Learning and Intelligent Systems Engineering (MLISE)* (pp. 267-271). IEEE.

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

Wright, R. T., Johnson, S. L., & Kitchens, B. (2023). Phishing Susceptibility in Context: A Multilevel Information Processing Perspective On Deception Detection. *MIS Quarterly*, *47*(2).

Wright, R., Marett, K., & Thatcher, J. (2014). Extending Ecommerce Deception Theory to Phishing.

Yu, S., Tack, J., Mo, S., Kim, H., Kim, J., Ha, J. W., & Shin, J. (2022). Generating videos with dynamics-aware implicit generative adversarial networks. *arXiv preprint arXiv:2202.10571*.