DEFENDING REALITY: HUMAN-AI COLLABORATION TO UNVEIL DEEPFAKE INFORMATION MANIPULATION

Shuyuan Mary Ho
Florida State University
School of Information
smho@fsu.edu

Marco Prandini
University of Bologna
Computer Science
marco.prandini@unibo.it

Franco Callegati
University of Bologna
Computer Science
franco.callegati@unibo.it

Shayok Chakraborty
Florida State University
Computer Science
schakraborty2@fsu.edu

Thomas Stephan Juzek
Florida State University
Computational Linguistics
tjuzek@fsu.edu

Yue Liu
Florida State University
School of Information
yl22t@fsu.edu

ABSTRACT

As deepfake information manipulation technology continues to evolve and propagate, its potential to mislead the public poses growing threat to societal trust. This paper outlines our research agenda exploring the role of explainable AI (XAI) in cyber defense and its effectiveness in safeguarding reality. Our study examines mechanisms for unveiling deepfakes in ways that enhance sensemaking and strengthen individual cyber defense self-efficacy in distinguishing authentic from manipulated information. To achieve this, we designed and simulated human-AI collaboration experiments with participants from the United States and Italy in Spring 2025. These experiments will generate paired datasets of real and deepfake artifacts across audio, graphic, visual and textual content. XAI—defined by the completeness and relevance of explanations regarding deepfake information—will be modeled based on the insights from the collaboration. Ultimately, this study contributes to social cybersecurity by empowering individuals and communities to recognize and defend against deepfake information manipulation.

Keywords

Deepfake, information manipulation, computer-mediated deception, human-AI collaboration, eXplainable AI, cyber defense self-efficacy

INTRODUCTION

Deepfake information manipulation, a form of compute-mediated deception [15-17], represents one of the most sophisticated challenges facing society today. This issue arises from the creative use and widespread adoption of rapidly evolving high-tech image manipulation software [7, 34]. A deepfake refers to a digital artifact—such as audio, graphic, visual, or textual content—originally created to convey authentic concepts but later manipulated convincingly using advanced technology. This specific form of manipulation involves digitally altering, modifying, or replacing images or videos to reshape perceptions of reality. Advanced editing tools (e.g., Adobe Photoshop, Adobe Premiere Pro, and Adobe Audition) have made it increasingly effortless to create and distribute deepfake disinformation. This proliferation of deepfakes poses a greater threat than disinformation in conversational exchanges. Unlike traditional misinformation, deepfakes leverage AI-driven techniques to manipulate images, audio, and video, generating highly misleading content. With minimal expertise, anyone can exploit this technology to fabricate and spread deceptive media.

A key challenge is that people often struggle to distinguish between genuine and manipulated content. Deepfake information can exist on a spectrum of truth and intent [18, 19]. In essence, while some deepfakes are deliberately created to knowingly spread falsehood, others are unknowingly shared by individuals unaware of their inauthenticity. The consequences of deepfake proliferation can be severe, affecting personal reputation and credibility—as seen in high-profile incidents like the Taylor Swift deepfake scandal [28]—and extending to broader societal domains, including public health, politics, and the economy. Recognizing artificially manipulated information is nearly impossible with the human eye alone. Individuals unfamiliar with these technologies may find it particularly difficult to differentiate between authentic and altered content. As a result, cyber citizens who lack the ability to detect deepfakes can inadvertently contribute to the spread of misinformation.

A pressing societal issue associated with deepfakes is their ability to influence public discourse. The harm caused by synthetic media extends beyond immediate falsehoods to long-term consequences, including the erosion of people's ability to distinguish truth from fiction. The widespread proliferation of deepfakes could trigger a crisis of confidence in audio and video recordings,

which have historically played a crucial role in holding individuals accountable [26]. As deepfakes become more convincing, viewers may begin to question the authenticity of even trustworthy imagery. The inability to judge digital information based on appearance alone increases the risk of an "infocalypse" [30], where mass distrust in audiovisual media—driven by deepfakes, misinformation, and manipulated content—has far-reaching consequences. In journalism, deepfakes can undermine credibility, erode public trust, and blur the line between fact and fiction. AI-generated news anchors and fabricated media [29] can be used to push political agendas, while journalists may hesitate to report on controversial topics for fear of inadvertently spreading fake news. In politics, deepfake videos can depict politicians making false statements or engaging in fabricated actions, potentially influencing elections and amplifying polarization to deepen political divisions. In the financial sector, the economy and stock market can be manipulated through fake videos announcing bankruptcies or financial crises. AI-generated media can also be weaponized to impersonate executives or mislead investors, as seen in the case of a deepfake video conference used to deceive a finance worker at a multinational firm in Hong Kong into transferring $25 million to fraudsters [5]. Beyond these domains, deepfakes threaten the very fabric of societal trust. In essence, our shared reality has been broken-down. When people begin to doubt all audiovisual evidence, it becomes increasingly difficult to establish facts. Courts and law enforcement agencies may struggle to verify video or audio evidence, complicating justice. Public consensus on critical issues such as public health and climate change may become harder to achieve, weakening social cohesion. Ultimately, if distrust in audiovisual media becomes widespread, misinformation can thrive, and institutional trust may erode, leading to a world where uncertainty dominates.

Given these challenges, the ability to differentiate authentic from manipulated information is crucial. This skill—deepfake information literacy—plays a vital role in cyber defense. Research suggests that individuals who are more informed about the impact of deepfakes tend to have higher self-efficacy in recognizing them [13]. However, people often overestimate their ability to detect deepfakes [20]. Enhancing both the ability and confidence to recognize deepfakes requires technology capable of explaining and differentiating manipulated content from authentic information. As such, this leads to our grand research question: *Can explainable AI improve individual's cyber defense self-efficacy in identifying deepfakes*? We hypothesize that XAI can enhance individuals' confidence and ability to recognize manipulated digital content. Our proposed deepfake research agenda consists of four phases: 1) ***Data generation***: we will generate a paired dataset of real and deepfake artifacts with a set of multimodal data sources (i.e., video, audio, image and text) that are currently unavailable in scientific communities. This will be achieved through a series of human-AI collaboration experiments. 2) ***Feature extraction***: Second, we will identify and extract features, clues and metadata unique to artificially generated content. 3) ***Model training***: using this paired dataset of real and deepfake artifacts, we will train a XAI model—with a multimodal data fusion model—that can explain deepfakes. The model will analyze and explain deepfake characteristics using complete and relevant information to enhance user' deepfake information literacy, engagement, and understanding. 4) ***User experience testing***: we will conduct user experience experiments to evaluate the effectiveness of XAI in improving cyber defense self-efficacy. We hypothesize that the more satisfied users are with AI explainability, the greater their confidence and ability to identify deepfakes. By enhancing deepfake information literacy and fostering trust in AI-assisted detection, our research aims to empower individuals with the explainable AI necessary to navigate an increasingly complex information landscape.

DEEPFAKE INFORMATION MANIPULATION

Computer-mediated deception refers to the intentional use of digital communication technologies—such as computers, smartphones, and the Internet—to mislead, manipulate, or deceive individuals or systems [15-17]. The emergence of deepfake AI technology has ushered in a new era of computer-mediated deception, where individuals can be misled not only by other people in interpersonal communication but also by AI-driven bots and programs designed to amplify deception. AI-mediated deception is particularly unpredictable due to the blended nature of AI-generated content. Generative AI can create highly realistic anthropomorphic conversations, images, and videos from textual instructions, making it difficult to discern fact from fiction. This evolving capability represents a paradigm shift in computer-mediated deception, introducing new challenges in detecting and mitigating misinformation [15-17]. Understanding the mechanisms behind these deceptive practices is crucial for improving the ability to identify false and misleading information.

Computer-mediated deception operates by interweaving falsehoods with reality, distorting people's perceptions of truth. The advancement of information technology and the Internet has propelled this issue to the forefront, making it an increasingly significant threat [15-17]. Alongside traditional news media, deceptive digital content has become a powerful tool for manipulating public perception. McCornack [22] first proposed the theory of information manipulation, arguing that deception extends beyond outright lies to include the strategic manipulation of information across different dimensions. His research examined how deceptive messages covertly violate conversational norms, influencing perceptions of deceptiveness and speaker competence, specifically including the manipulations of quantity, quality, relation, relevance, and manner of information [23]. These findings highlight how various forms of information distortion shape public perception, emphasizing the need for robust detection strategies in the digital age.

As electronic devices become increasingly integrated into daily life, computer-mediated deception continues to expand in multiple directions. This form of deceptive behavior includes social engineering tactics, such as online chats designed to manipulate recipients into divulging personal information, as well as the spread of misleading content intended to damage reputations. Research has identified significant differences between truthful and deceptive messages, particularly in message content [35]. Deceptive messages tend to be richer in quantity, variety, sentiment, and specificity. However, no significant difference in formality have been observed between truthful and deceptive messages [35]. Deceivers' messages often exhibit higher levels of affective language and a greater use of words, verbs, modifiers, and noun phrases compared to those from truth-tellers [35]. Despite these patterns, message recipients generally struggle to distinguish between true and false information, especially in informal contexts [4].

The rise of deepfake AI technology has introduced a broad spectrum of applications, ranging from facial recognition and speech synthesis to image enhancement, voice conversation, video interpretation, and translation. AI-driven technology can alter or swap facial features in videos and images, creating media that appear highly authentic [34]. These technologies enable the manipulation of facial expressions, vocal nuances, and gestures, making it possible to fabricate realistic portrayals of individuals engaging in actions or statements that never actually happened [21]. Such deepfake media can be deceptively used as fabricated evidence in court cases or false insurance claims. The ease with which AI can easily modify and convert digital content underscores the growing challenge of detecting and mitigating manipulated media [21].

EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI)

The growing proliferation of deepfake media amplifies the urgent need for reliable detection methods. While certain applications, such as Microsoft Azure AI Vision [25] and Face++ [24], can verify identities through facial recognition, emotion analysis, and body detection, they primarily focus on face matching rather than detecting deepfake manipulation. These technologies identify visual irregularities but do not assess the authenticity of images, making them insufficient for real-world deepfake detection. Existing detection methods and benchmarks remain inadequate in addressing the evolving sophistication of deepfake technologies.

AI is often perceived as a black-box technology by both developers and users. To enhance transparency and trust, eXplainable AI (XAI) has been proposed as a means to improve AI adoption [27]. By definition, XAI refers to AI models that can explain or interpret their decisions, in an intelligible and meaningful way for a targeted audience [2, 8]. This means AI-generated explanations should provide users with insights into how the systems reaches its conclusions. Explainability helps uses understand these insights, while interpretability enables them to make sense of the information provided [27]. Saeed and Omlin [27] identified five key perspectives of XAI—end-user, developer, industrial, scientific and regulatory—each shaped by the unique needs of different audiences. To address societal challenges in cyber defense against deepfakes, our research examines AI explainability, particularly its role in influencing users' emotional engagement and fostering confidence in cyber defense self-efficacy. By enhancing user satisfaction with XAI-driven image analysis engines, we aim to empower individuals to recognize and counter deepfake manipulation.

XAI has the potential to bridge the gap between truth and deception by providing users with the tools to discern manipulated media. While scientific advancements demonstrate the positive adoption of AI technology, addressing the dark side of methodological, societal, and technological challenges of AI adoption requires collective wisdom and joint efforts [11]. Building a "Good AI Society" depends on designing AI systems that enhance individual self-efficacy, support autonomous decision-making, and amplify human capabilities—without diminishing human responsibility [9, 10]. A pressing question remains as to whether these deepfake detection systems and explainable AI can effectively equip online users and society with the confidence and capability to identify deepfakes.

CYBER DEFENSE SELF EFFICACY

Self-efficacy refers to individuals' beliefs in their ability to successfully perform certain tasks. Bandura [3] suggested that self-efficacy influences one's choices of activities, the effort they put in, and their persistence in the face of challenges. Self-efficacy can be shaped by various factors, including personal experiences, vicarious learning, verbal persuasion and encouragement, and emotional arousal. People are more likely to avoid tasks that they perceive as beyond their abilities but will engage in activities they feel confident in handling. Higher self-efficacy increases the likelihood of succeeding at tasks, contributing to the development of mastery experience. Individuals can gain self-efficacy not only through direct experience but also by observing others' successes, which reinforces their belief that they too can achieve similar outcomes. Nonetheless, differences in past experiences lead to varying levels of generalized self-efficacy expectations [32]. Self-efficacy in technology adoption evolves over time, from initial use to continued engagement. Wang, Harris, et al. [33] found that users' self-efficacy has a strong influence on the initial adoption of technology, but its impact diminishes as they become more accustomed with the

technology. Over time, users' satisfaction with technology becomes a key factor in promoting continued use and habit formation.

Cyber defense refers to strategies, practices and measures employed to protect information assets. As Schneier [31] highlighted, "amateurs hack machines; professionals hack people." Information can be manipulated through deception, such as deepfake technology, which alters reality and hacks people's mind. People's cyber defense self-efficacy in seeing through deepfakes can be more directly boosted through the use of explainable AI (XAI) systems as priming stimuli. Consciousness of cyber defense [14]—as a state of cyber situational awareness—the ability to perceive and think critically during a cyber threat serves as the foundation for cyber defense self-efficacy. For example, self-efficacy in technology adoption is commonly reflected in behaviors like online banking, shopping, remote work, and distance learning. Self-efficacy in cybersecurity—such as recognizing phishing emails and managing passwords—has been studied extensively. However, self-efficacy in identifying deepfakes has been less explored and is often linked to personality traits [1]. While deepfake training programs can help improve detection accuracy, they can also increase emotional distress and reduce self-efficacy [6]. Cyber defense self-efficacy specifically refers to one's belief in their ability to distinguish deepfakes from genuine content in order to protect their personal beliefs of reality. An effective approach to improving self-efficacy and "seeing through" deepfakes requires consistent exposure. As with many skills, success breeds confidence. Facilitating user experience with deepfake detection through explanatory exposure can help users feel more confident and effective in their ability to identify manipulated content. Based on our observations, users who are guided through the process of recognizing deepfakes tend to develop greater self-confidence. Therefore, we hypothesize that user satisfaction with AI explainability can positively influence cyber defense self-efficacy.

User satisfaction plays a critical role in the success and effectiveness of technology. Satisfied users are more likely to use technology effectively, leading to improvements in their performance and productivity. They are also more inclined to perceive technology as useful and easy to use, which encourages positive attitudes and intention to continue using it [12]. When users are satisfied with explainable AI, it indicates that the system meets their needs and expectations [12]. XAI system can encourage users to engage more deeply with their goals, gaining insights that enhance their understanding. Higher satisfaction increases individuals' sense of cyber defense self-efficacy, making them feel more capable of protecting themselves and motivating them to explore and enhance their potentialities. In this context, user satisfaction plays a pivotal role in bolstering cyber defense self-efficacy in recognizing and defending against deepfakes.

METHOD: HUMAN-AI COLLABORATION EXPERIMENT DESIGN

The research team will collect multimodal data from multimodal sources through a human-AI collaboration experiment, which will be conducted as a campus-wide activity during Spring 2025. The research protocol has been reviewed and approved by the FSU Human Subjects Committee. Approximately 100 teams, totaling 300 participants from the United States and Italy, will be recruited for the study. In phase 1, participant demographics will be collected following written informed consent. A *pre*-test and *post*-test-survey will be designed to capture participants' perceptions and experiences. The research protocol will enable participants to interact with various AI technologies on a daily basis. Participants will receive task assignments and will be asked to generate conversations with AI and maintain diaries documenting their interactions. Conversations and interactions between participants and AI technologies will be collected over a period of approximately 5 days. In phase 2, participants will be tasked with using modification techniques to generate their own deepfakes videos, which will be paired with the corresponding ground truth. The data collection period may be extended, if necessary. The study aims to collect 1,000 files of ground truth, paired with 1,000 files of deepfake artifacts. The dataset will also include multimodality of interaction data from participants. In phase 3, participants will be randomly assigned files—generated by other participants in the same study—to evaluate and judge whether they are authentic or manipulated. Additionally, 10 focus group sessions will be conducted to interview team-based participants about their experiences creating both ground truth and deepfake manipulated information.

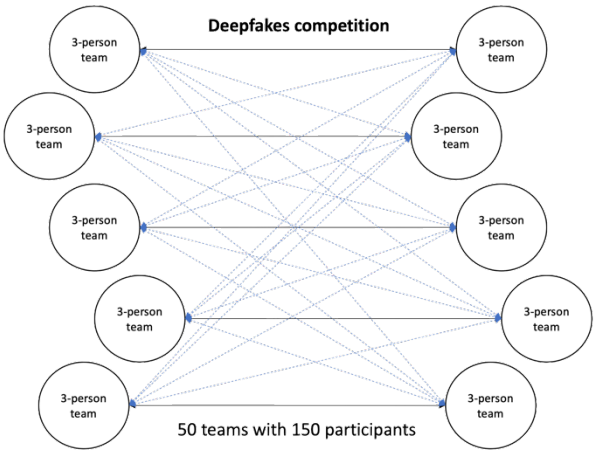

Figure 1. Research experiment design.

RESEARCH SIGNIFICANCE, IMPLICATION, AND CONTRIBUTION

The research will collect multimodal data from the same source, creating a unique dataset that does not currently exist in the scientific community. Developing this multimodal dataset is crucial to advancing the fields of information science, computer science, and linguistics, furthering AI research, and enhancing human capacity. All data will be made publicly available,

amplifying the broader impact of the study. The deepfake research problem is foundational to core areas of information science, computer science, as well as linguistics. The dataset will be analyzed and modeled to contribute both scientifically and theoretically to these fields. Additionally, the research will have practical implications for defending society by advancing knowledge and providing a scientific approach to verifying deepfake content.

This research agenda will first produce a set of curricula and tools aimed at increasing deepfake detection skills across the public. Second, the project will raise societal awareness of both the positive (e.g., information sharing, connection, support) and negative (e.g., deepfakes, misinformation, disinformation) consequences associated with social technology and media. Third, the project will enhance individual and collective situational awareness and efficacy in identifying deepfake manipulation. Fourth, strategies and opportunities will be devised to shift public perceptions of mis/dis-information and deepfakes, helping to combat societal mistrust.

CONCLUSION

Deepfake technology will continue to exacerbate public confusion by spreading misleading information. People often trust the information shared within their circles of friends and family because determining the authenticity of content—whether photos, images, graphs or even audio and videos—can be challenging. Our study highlights the importance of developing effective data fusion and analyzing systems to detect deepfake content. As deepfake technology becomes more accessible, these analysis systems can help individuals navigate the confusion. This not only enables more effective identification of false information but also reduces the likelihood of spreading misinformation, contributing to a safer online community. XAI-powered deepfake analysis addresses the gap in user cyber defense self-efficacy when recognizing manipulated information, significantly boosting user confidence in detecting deepfakes. With AI explainability and a supportive user interface, user's cyber defense self-efficacy can be further strengthened. Future research should explore these dynamics in greater depth to ensure that users are adequately equipped to navigate the ever-evolving digital landscape.

ACKNOWLEDGMENTS

The first author is a Fulbright Scholar for the 2024-2025 Lectureship in Big Data, Networking and Artificial Intelligence at the University of Bologna. The authors acknowledge the support of the U.S.–Italy Fulbright Commission, Institute of International Education (IIE), as well as the support of the Florida State University Council on Research and Creativity (CRC) SEED grant (award #047041, 5/15/2024—12/30/2025). Additional support was provided by Project SERICS (PE00000014) under the MUR National Recovery and Resilience Plan, funded by the European Union—NextGenerationEU.

REFERENCES

1. Abraham, J., H.A. Putra, T. Prayoga, H.L.H.S. Warnars, R.H. Manurung, and T. Nainggolan, *Predictio of self-efficacy in recognizing deepfakes based on personality traits.* F1000 Research, 2023. **11**: pp. 1529 (1-18). doi:10.12688/f1000research.128915.3.
2. Arrieta, A.B., N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, *Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI.* Information Fusion, 2020. **58**: pp. 82-115. doi:10.1016/j.inffus.2019.12.012.
3. Bandura, A., *Self-efficacy: Toward a unifying theory of behavioral change.* Psychological Review, 1977. **84**(2): pp. 191-215. doi:10.1037/0033-295X.84.2.191.
4. Burgoon, J.K., J.P. Blair, T. Qin, and J.F. Nunamaker, *Detecting deception through linguistic analysis.* Intelligence and Security Informatics, 2003. **2665**: pp. 91-101.
5. Chen, H. and K. Magramo, Finance worker pays out $25 million after video call with deepfake 'chief financial officer', in CNN, on February 4, 2024. Available from: https://www.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html.
6. Diel, A., A. Bäuerle, and M. Teufel, Inability to detect deepfakes: Deepfake detection training improves detection accuracy, but increases emotional distress and reuces self-efficacy, in Social Science Research Network (SSRN), 2024. Retrieved from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5002585.
7. Dolhansky, B., J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C.C. Ferrer, The deepfake detection challenge (DFDC) dataset, in Meta AI, arXiv:2006.07397. 2020. Retrieved from: https://ai.meta.com/datasets/dfdc/.
8. Došilović, F.K., M. Brčić, and N. Hlupić. *Explainable artificial intelligence: A survey*. in *Proceedings of the 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. 2018. Opatija, Croatia: IEEE. doi:10.23919/MIPRO.2018.8400040.
9. Floridi, L., J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, and E. Vayena, *AI4People—An ethical framework for a good AI society: Opportunities, risks, principles and recommendations.* Minds and Machines, 2018. **28**: pp. 689-707. doi:10.1007/s11023-018-9482-5.

10. Floridi, L. and J.W. Sanders, *On the morality of artificial agents.* Minds and Machines, 2004. **14**: pp. 349-379. doi:10.1023/B:MIND.0000035461.63578.9d.
11. Hagendorff, T. and K. Wezel, *15 challenges for AI: or what AI (currently) can't do.* AI & Society, 2020. **35**: pp. 355-365. doi:10.1007/s00146-019-00886-y.
12. Henry, J.W. and R.W. Stone, *Computer self-efficacy and outcome expectancy: The effects on the end-user's job satisfaction.* Computer Personnel, 1995. **16**(4): pp. 15-34. doi:10.1145/219716.21972.
13. Ho, S.M., J. D'Arcy, and Y. Liu, *Fooling the eye, convincing the mind: User experience with explainable AI in seeing through deepfakes.* under review.
14. Ho, S.M. and M. Gross, *Consciousness of cyber defense: A collective activity system for developing organizational cyber awareness.* Computers & Security, 2021. **108**: pp. 102357 (1-18). doi:10.1016/j.cose.2021.102357.
15. Ho, S.M. and J.T. Hancock. *Computer-mediated deception: Collective language-action cues as stigmergic signals for computational intelligence*. in *Proceedings of the 2018 51th Hawaii International Conference on System Sciences (HICSS-51)*. 2018. Big Island, Hawaii: University of Hawaii. doi:hdl.handle.net/10125/50098.
16. Ho, S.M. and J.T. Hancock, *Context in a bottle: Language-action cues in spontaneous computer-mediated decepion.* Computers in Human Behavior, 2019. **91**: pp. 33-41. doi:10.1016/j.chb.2018.09.008.
17. Ho, S.M., J.T. Hancock, C. Booth, and X. Liu, *Computer-mediated deception: Strategies revealed by language-action cues in spontaneous communication.* Journal of Management Information Systems, 2016. **33**(2): pp. 393-420. doi:10.1080/07421222.2016.1205924.
18. Karlova, N.A. and K.E. Fisher, *A social diffusion model of misinformation and disinformation for understanding human information behavior.* Information Research, 2013. **18**(1).
19. Karlova, N.A. and J.H. Lee. *Notes from the underground city of disinformation: A conceptual investigation*. in *Proceedings of the American Society for Information Science and Technology*. 2011. ASIS&T. doi:10.1002/meet.2011.14504801133.
20. Kobis, N.C., B. Dolezalova, and I. Soraperra, *Fooled twice: People cannot detect deepfakes but think they can.* iScience, 2021. **24**(11): pp. 103364. doi:10.1016/j.isci.2021.103364.
21. Maras, M.-H. and A. 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, 2019. **23**(3): pp. 255-262. doi:10.1177/1365712718807226.
22. McCornack, S.A., *Information manipulation theory.* Communication Monographs, 1992. **59**(1): pp. 1-16. doi:10.1080/03637759209376245.
23. McCornack, S.A., T.R. Levine, K.A. Solowczuk, H.I. Torres, and D.M. Campbell, *When the alteration of information is viewed as deception: An empirical test of information manipulation theory. Communications Monographs.* Communications Monographs, 1992. **59**(1): pp. 17-29. doi:10.1080/03637759209376246.
24. MegVII. Face++. 2024. Available from: https://www.faceplusplus.com/.
25. Microsoft. Azure AI Vision. 2024. Available from: https://azure.microsoft.com/en-us/products/ai-services/ai-vision.
26. Rini, R., *Deepfakes and the epistemic backstop.* Philosophers' Imprint, 2020. **20**(24): pp. 1-16.
27. Saeed, W. and C. Omlin, *Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities.* Knowledge-based Systems, 2023. **263**: pp. 110273: 1-24. doi:10.1016/j.knosys.2023.110273.
28. Saner, E., Inside the Taylor Swift deepfake scandal: 'It's men telling a powerful woman to get back in her box', in The Guardian, Online, on Jan 31, 2024. Available from: https://www.theguardian.com/technology/2024/jan/31/inside-the-taylor-swift-deepfake-scandal-its-men-telling-a-powerful-woman-to-get-back-in-her-box.
29. Satariano, A. and P. Mozur, The people onscreen are fake. The disinformation is real., in The New York Time, on February 7. Available from: https://www.nytimes.com/2023/02/07/technology/artificial-intelligence-training-deepfake.html.
30. Schick, N., *Deepfakes and the infocalypse: What you urgently need to know*. 2020: Monoray.
31. Schneier, B. *Schneier's blog on security*. 2013 [cited 2024 April]; Available from: https://www.schneier.com/blog/archives/2013/03/phishing_has_go.html.
32. Sherer, M., J.E. Maddux, B. Mercandante, S. Prentice-Dunn, B. Jacobs, and R.W. Rogers, *The self-efficacy scale: construction and validation.* Psychological Reports, 1982. **51**(2): pp. 663-671. doi:10.2466/pr0.1982.51.2.663.
33. Wang, C., J. Harris, and P. Patterson, *The roless of habit, self-efficacy, and satisfaction in driving continued use of self-service technologies: A longitudinal study.* Journal of Service Research, 2013. **16**(3): pp. 400-414. doi:10.1177/1094670512473200.
34. Westerlund, M., *The emergence of deepfake technology: A review.* Technology Innovation Management Review, 2019. **9**(11): pp. 39-52. doi:10.22215/timreview/1282.
35. Zhou, L., D.P. Twitchell, T. Qin, J.K. Burgoon, and J.F. Nunamaker Jr. *An exploratory study into deception detection in text-based computer-mediated communication*. in *Proceedings of the 2003 Hawaii International Conference on System Sciences (HICSS-36)*. 2003. Big Island, Hawaii: IEEE. doi:10.1109/HICSS.2003.1173793.