Undergraduate Thesis Prospectus

Improving Deepfake Detection Systems in Financial Institutions

(technical research project in Systems Engineering)

Competing Responses to the Growing Threat of Deepfakes

(sociotechnical research project)

by

Fahima Mysha

November 8, 2024

technical project collaborators:

Padma Lim, Rhea Agarwal, Baani Kaur, Vishnu Lakshmanan, Drake Ferri

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

Fahima Mysha

Technical advisor: Gregory Gerling, Department of Systems Engineering

STS advisor: Peter Norton, Department of Engineering and Society

General Research Problem

How can the use of emerging technologies be managed to balance innovation with ethics?

Artificial intelligence, machine learning and other developments in digital technology can support valuable innovations. Yet the rapid development of AI has outpaced regulation, threatening data privacy, security, user trust, and authenticity. Criminals have used an AI deepfake to impersonate a CFO in a $25 million scam, and highly automated robotic vehicles have caused deaths (Chen and Magramo, 2024).

Improving Deepfake Detection Systems in Financial Institutions

How can deepfake detection systems be improved to address vulnerabilities in financial systems related to spoofed voice authentication, by identifying the key differences between deepfake and real voices and understanding the factors that affect their detection by both humans and banking systems?

The capstone project, advised by Dr. Gregory Gerling in the Systems Engineering department, explores the darker side of deepfakes and their criminal potential, with team members Padma Lim, Rhea Agarwal, Baani Kaur, Vishnu Lakshmanan, and Drake Ferri. The goal is to bypass voice authentication systems while identifying the specific combination of factors that influence the likelihood of detection. This includes understanding how different factors, such as the length of the training sample or the volume, affect the detection of cloned voices by humans, Automatic Speaker Verification (ASV) systems, and bank security systems. The technical project aims to uncover weaknesses in current financial security systems that are caused by AI-generated deepfake voices.

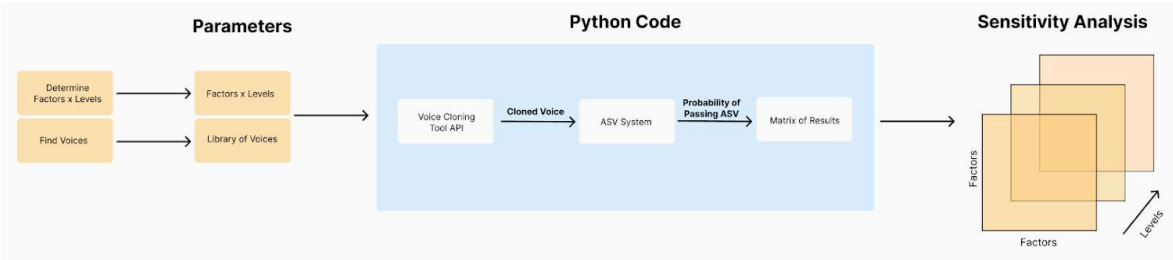

Figure 1. This figure, created by Padma Lim, Rhea Agarwal, Baani Kaur, Vishnu Lakshmanan, Drake Ferri, and Fahima Mysha as part of the technical capstone project, illustrates the experimental model to assess vulnerabilities.

As shown in Figure 1, the initial steps involve identifying factors that can be adjusted across different levels. For example, one key factor is the length of the training sample, which is categorized into three distinct levels: 15 seconds, 60 seconds, and 300 seconds. Initially, the voices will be from members within the group, with the possibility of expanding the library of voices using GitHub. Using both commercial and open-source voice cloning tools, a library of cloned voices will be created. To assess the effectiveness of the cloned voices, a liveness check will be performed by calling the bank's customer service line. The cloned voices will be transmitted via a VoIP connection to test whether the bank's system can correctly identify them as live or synthetic. A human-subjects experiment will be conducted to determine if people can detect if the voices they hear are cloned voices, and if the voices are similar to the intended victim. An Automatic Speaker Verification (ASV) system will test the voice clones' ability to bypass the security system. The ASV results will include a matrix showing the probability of successfully bypassing the system and not being detected as a clone. A sensitivity analysis will identify the combination of factors that differentiate a real voice from a cloned voice. Python software will integrate the various components, and specifically for matrix creation and sensitivity analysis. Following testing with ASVs, there are plans to test directly with bank security systems to assess the effectiveness of the cloned voices in real-world scenarios. This

work will not only highlight vulnerabilities in current voice authentication methods, but also give insights for developing improved security measures.

Competing Responses to the Growing Threat of Deepfakes

In the U.S., how have financial institutions, social media platforms, news media, and government agencies responded to the growing threat of deepfakes?

The proliferation of deepfakes is a threat to financial institutions, social media platforms, news media, and government agencies. To protect themselves they must adapt to rapidly developing technology. Deepfakes can defeat financial institutions' authentication systems, exposing them to substantial material and reputational risks (Bansal et al., 2023). Ileana van der Linde, Executive Director at JP Morgan, highlights risks such as extortion scams impersonating family members, government officials, and bank employees. She noted that fraud scams led to losses of over $12.5 billion in 2023, with a 22% increase from the previous year (van der Linde, 2024). KPMG emphasizes the need for a governance framework that "embraces disruptive technologies and encourages innovation while ensuring risks are identified and managed" for organizations to thrive digitally (KPMG, 2023). This highlights the interconnectedness of adopting new technologies while also developing solutions to mitigate associated risks.

Social media platforms have actively tried to limit the spread of misinformation and fake news, while maintaining the freedom of creativity and expression. Facebook launched The Deepfake Detection Challenge (DFDC) to detect malicious deepfakes. Over 2,000 participants contributed, resulting in over 35,000 detection models (Schroepfer, 2020). Building on this, Meta now flags AI-altered content across its platforms with labels, balancing the need to counteract advanced AI threats with the commitment to free expression (Meta, 2024). Social media

platforms and news media have collaborated to address this challenge jointly. Reuters partnered with Facebook to fact-check content during the 2020 U.S. presidential election, with Facebook responding to calls for greater accountability in removing fake news (Reuters, 2020). Similarly, The Wall Street Journal mobilized a team of 21 journalists to detect misinformation, including deepfakes, in the lead-up to the election (Digiday, 2019).

In 2019, the House Intelligence Committee convened to assess “the national security threats posed by AI-enabled fake content,” focusing on detection and mitigation strategies as well as exploring the roles of the public, private, and societal sectors (HPSCI, 2019). This meeting led to DARPA's creation of the Semantic Forensics (SemaFor) program to combat deepfakes and manipulated media. According to DARPA, "...research investments in detecting, attributing, and characterizing manipulated and synthesized media, known as deepfakes, have resulted in hundreds of analytics and methods" (DARPA, 2024). The House Intelligence Committee also helped pass the Deepfake Report Act of 2019, requiring the Secretary of Homeland Security to produce an annual report on deepfake technology use (US Senate, 2019).

The RAND Corporation, a U.S.-funded think tank, highlights the escalating threat of deepfakes, stressing the need for strategic countermeasures to prevent AI-generated disinformation from undermining journalism, eroding trust, polarizing society, and manipulating elections (Helmus, 2022). RAND urges policymakers to address these risks proactively by implementing robust detection methods and policy frameworks to counter the destabilizing effects of deepfakes. In a recent publication regarding the 2024 U.S. Election, RAND further stresses that federal agencies should work closely with technology and cybersecurity experts to reveal deepfakes “... amplify misleading messages and sow discord” (Posard et al, 2024).

Published Research

The financial sector is a significant and valuable digital target for criminals. Financial institutions have become increasingly aware of the escalating threat posed by deepfakes, which expose vulnerabilities and heighten the risk of fraud. Commercial and open source voice cloning tools, now widely available, bypass traditional security like voice ID. “As of March 2023, a total of 47 open-source voice cloning software programs can be found on GitHub…” (Bansal et al., 2023).

Yadlin-Segal and Oppenheim (2021) argue that deepfake technologies pose serious societal risks, especially for vulnerable groups like women and children, due to the potential for abuse and harassment. They emphasize that journalists see themselves as sociopolitical “gatekeepers” whose role remains crucial even in the face of machine-generated content, stating that "gatekeepers' authority to police content doesn't evaporate just because it was made by machine learning" (Yadlin-Segal and Oppenheim, 2021). With disinformation on the rise, journalists aim not only to inform the public but also to advocate for social media regulation to protect marginalized groups and maintain societal trust. Deepfakes present a challenge for news media, prompting journalists to adopt stronger fact-checking methods. Vizoso et al. (2021) note that “while the media is focused on training journalists for its detection, online platforms tended to fund research projects whose objective is to develop or improve media forensics tools.” This highlights the ongoing collaborative efforts across industries, with both journalists and online platforms working together to combat deepfakes and misinformation.

Social media platforms have revolutionized news accessibility, making it easier for the public to stay informed. However, this increased accessibility has also led to a rise in the spread of rumors, fake news, and disinformation (Guess et al., 2019). Social media platforms are deeply

invested in combating the growing threat of deepfakes. Facebook has launched the Deepfake Detection Challenge, in collaboration with major tech companies like Amazon and Microsoft, to combat the threats posed by deepfakes (Strickland, 2020).

Governments are becoming more aware of the dual nature of deepfake technology, as it provides new opportunities for innovative public communication while also posing significant risks to democratic stability. "Deepfake propaganda and election meddling, and the disinformation they seed, threaten efficient governance for all democracies, if not democracy itself" (Kietzmann et al., 2020). Governments are increasingly recognizing the potential dangers of deepfakes, particularly in the context of elections, and are stepping in with legislation to combat this issue. Texas and California have passed laws criminalizing the publishing and distribution of deepfake videos intended to harm political candidates or deceive voters during elections (Castro, 2020).

References

Bansal, N., Boissard, M., Garrido-Lecca, M., Han, X., Munts, M., Noboa, A., Reinfeldt, G., & Shi, Y. (2023). Deepfakes and disinformation in the finance sector: Strategies to prevent and deter. Columbia SIPA and Bank of America. https://www.sipa.columbia.edu/sites/default/files/2023-05/For%20Publication_BOfA_PollardCartier.pdf

Castro, D. (2020). Deepfakes are on the rise: How should the government respond? Government Technology. https://www.govtech.com/policy/Deepfakes-Are-on-the-Rise-How-ShouldGovernment-Respond.html

Chen, H., & Magramo, K. (2024, February 4). *Finance worker pays out $25 million after video call with Deepfake "chief financial officer."* CNN. https://www.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html

DARPA (2024, March 14). Defense Advanced Research Projects Agency. DARPA's efforts in countering synthetic media threats.

Digiday. (2019, July 1). The Wall Street Journal has 21 people detecting deepfakes. https://digiday.com/media/the-wall-street-journal-has-21-people-detecting-deepfakes/

Guess, A., Nagler, J., & Tucker, J. (2019, January 9). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. https://www.science.org/doi/10.1126/sciadv.aau4586

Helmus, T. C. (2022, July). Artificial Intelligence, deepfakes, and disinformation. https://www.rand.org/content/dam/rand/pubs/perspectives/PEA1000/PEA1043-1/RAND_PEA1043-1.pdf

HPSCI (2019). U.S. House Permanent Select Committee on Intelligence. HPSCI Democrats release report on the risks of AI and deepfakes in the context of national security. Democrats on the House Intelligence Committee.

Kietzmann, J., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. (2020). Deepfakes: Trick or treat? *Business Horizons*, *63*(2), 135–146. https://doi.org/10.1016/j.bushor.2019.11.006

KPMG. (2023, November 1). Emerging technology governance: A new era of strategic risk management. KPMG. https://kpmg.com/us/en/articles/2023/emerging-technology-governance.html

Meta. (2024, April). Meta's approach to labeling AI-generated content and manipulated media. Meta. https://about.fb.com/news/2024/04/metas-approach-to-labeling-ai-generated-content-and-manipulated-media/

Posard, M. N., Helmus, T. C., Woods, M., & Chandra, B. (2024, April). The 2024 U.S. election, Trust, and Technology. https://www.rand.org/content/dam/rand/pubs/perspectives/PEA3000/PEA3073-1/RAND_PEA3073-1.pdf

Reuters (2020, Feb. 12). Facebook starts fact-checking partnership with Reuters. https://www.reuters.com/article/us-facebook-partnership-reuters/facebook-starts-fact-checking-partnership-with-reuters-idUSKBN2062K4/

Schroepfer, M. (2020, September 9). *Think like the bad guys: An interview with Cristian Canton Ferrer, AI Red Team lead*. Tech at Meta. https://tech.facebook.com/ideas/2020/9/dfdc/

Strickland, E. (2020a). Facebook takes on deepfakes. *IEEE Spectrum*, *57*(1), 40–57. https://doi.org/10.1109/mspec.2020.8946309

US Senate (2019, Oct. 29). S.2065: Deepfake Report Act of 2019. 116th Congress. https://www.congress.gov/bill/116th-congress/senate-bill/2065/text

van der Linde, I. (2024, July 9). *Ai scams, deep fakes, impersonations ... oh my: J.P. Morgan*. AI Scams, Deep Fakes, Impersonations ... Oh My | J.P. Morgan. https://www.jpmorgan.com/insights/fraud/fraud-protection/ai-scams-deep-fakes-impersonations-oh-my

Vizoso, Á., Vaz-Álvarez, M., & López-García, X. (2021, March 3). Fighting deepfakes: Media and internet Giants' converging and diverging strategies against hi-tech misinformation: Article. https://www.cogitatiopress.com/mediaandcommunication/article/view/3494

Yadlin-Segal, A., & Oppenheim, Y. (2021). Whose dystopia is it anyway? Deepfakes and social media regulation. Convergence, 27(1), 36-51. https://doi.org/10.1177/1354856520923963