**Designing and Implementing a Scalable Data Loss Prevention System: A Full-Stack Approach**

**The Ethics of AI: Addressing Bias in Machine Learning Systems**

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Natalie Yee

November 8, 2024

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.

ADVISORS

Caitlin D. Wylie, Department of Engineering and Society

Brianna Morrison, EN-Comp Science Dept

## Introduction

Data is fundamental to modern society and powers many sectors such as finance and healthcare. However, there are rising concerns about the lack of data privacy and security as we increasingly rely on technologies that collect, store, and analyze large amounts of personal data. Mishandling personal data not only leads to financial losses and identity theft but also erodes user trust, as shown by the 2018 LastPass breach that left many users upset and distrustful. During this breach, attackers had copied a backup of customer vault data, which included unencrypted website URLs, usernames, and passwords (Sugunaraj, 2024). In addition to how data is protected, a problem is how data is being used to train Artificial Intelligence (AI) models. Instances of bias in AI models based on the datasets have occurred in hiring models. For example, Amazon's recruitment algorithm was found to favor male candidates over women, as it was trained on historical hiring data reflecting past gender biases, leading to inequitable hiring outcomes (Dastin, 2022). My technical research will focus on building secure, scalable systems to protect sensitive data, while my STS research will focus on what contributes to bias in AI. These papers will explore both the technical solutions and the broader societal challenges of data and how it is used.

## Designing and Implementing a Scalable Data Loss Prevention System: A Full-Stack Approach

The technical topic my research will focus on is the critical problem of securing sensitive data against cyberattacks and data breaches. Cyberattacks are particularly damaging because they can compromise sensitive information, disrupt operations, and break trust in digital systems. With the increased frequency and sophistication of cyberattacks, it is more important than ever that secure systems are in place to protect data. It is common knowledge among engineers in the

cybersecurity field that strong encryption, secure communication protocols, and access controls are essential to safeguarding data (Sun et al., 2014). However, as technology evolves to include cloud computing and AI, data security methods need to also evolve to cover these additional technologies. Cloud technologies must be secure because they store data virtually and can scale with increasing datasets. I plan to investigate how we can design and implement secure cloud architectures that both ensure data protection but also are scalable for growing datasets.

Under the guidance of a cybersecurity professor, my research will include real-world case studies, simulations, and prototype development to evaluate the effectiveness of different security models. Case studies will be taken from documented data breaches and security incidents in industries that rely on cloud computing such as finance and healthcare. By analyzing past incidents, I will be able to identify common vulnerabilities and the effectiveness of existing security measures. Simulations will be used to model network environments and test how security models respond to different cybersecurity attacks. These simulations will allow me to understand how security models respond to real-time attacks. One area my research will focus on is the use of AI in detecting and preventing cyberattacks in real-time. Engineers know that machine learning models can be trained to recognize suspicious patterns in network traffic (Alshammari & Aldribi, 2021), but there is still a lot of information to be learned about how these models can be optimized for large-scale environments. Alshammari and Aldribi's (2021) paper helps inform the technical side of my research by providing evidence that machine learning-based models are effective at scaling protection against cyber threats. Sun and coauthors' paper is credible because it was peer reviewed and published in the International

Journal of Distributed Sensor Networks. Sun and coauthors' (2014) paper informed me of the importance of combining policy and technical measures to strengthen overall system security.

While AI offers powerful capabilities for cybersecurity, it's important to acknowledge that these systems may have inherent limitations and potential biases. AI models can inherit biases from training data or develop blind spots that attackers could exploit. My research will examine these limitations and implement safeguards to ensure reliable and unbiased security detection.

Through my research, I hope to contribute to this field by developing a prototype system that integrates AI for enhanced data security. I will test the prototype by conducting simulated attacks on the system. This prototype could potentially be applied to cloud environments to offer organizations an active approach to data security.

**The Ethics of AI: Addressing Bias in Machine Learning Systems**

The problem that my STS research will address is bias in AI systems, specifically within hiring processes, where biased AI decisions can amplify existing societal inequalities. It's important to study ethical AI use in hiring because it represents a critical gateway to economic opportunity, where algorithmic bias can have lasting impacts on individuals' careers. These biases can result in discriminatory hiring practices which impact the diversity of organizations. These biases can be caused by the datasets used to train AI models or by the algorithms themselves (Parikh et al., 2019). For instance, AI tools used in hiring have been shown to favor male candidates over equally qualified women, as these algorithms have learned from historical data that reflect societal biases (Vasconcelos et al., 2018).

From Kaur and coauthors' (2022) paper, I learned that building trustworthy AI requires fairness, accountability, and transparency in design. These principles form the foundation of my ethical frameworks and are rooted in deontological ethics, which emphasize duty and adherence to moral rules. My STS research aims to assess how hiring algorithms can meet ethical standards using these frameworks. Huang and coauthors' (2023) main argument in their paper is that without consideration of ethical concerns, AI systems can cause bias. Their overview of AI ethics provides a broader, consequentialist perspective, highlighting the societal impacts of biases. This information will be integrated to support my argument that fair AI development is needed to prevent societal harm.

From Parikh and coauthors' (2019) paper, I learned how bias in healthcare AI can worsen existing disparities, emphasizing the need of fairness to prevent societal harm. This aligns with justice-based ethics, which stresses equitable treatment and the avoidance of harm. I will use justice-based ethics to evaluate how hiring algorithms impact different demographic groups and whether they promote or hinder equitable opportunities. Dastin (2022) conducts a retrospective case study on AI bias with real-world implications, with a specific focus on observed failures, whereas Vasconcelos and coauthors (2018) propose a more theoretical approach, modeling principles for mitigating bias. I will use Dastin's work to inform my STS research on outcomes, while Vasconcelos and coauthors' (2018) framework will be used to inform the theoretical foundation of my work.

Nadeem and coauthors (2020) identify common causes of bias to include lack of diversity in training data and developers. One of the main sources of evidence that this paper references is statistics from UNESCO showing low representation of women in AI-related jobs, as well as examples of biased outcomes in recruitment software. They argue that gender bias is a significant ethical concern and that a more comprehensive approach is needed to mitigate gender bias. I will use their framework to analyze evidence of gender bias in hiring algorithms, specifically by investigating training datasets and development teams in my case studies.

Duncan and Mcculloh (2024) is a reliable source because it went through a rigorous peer review process to be published in the proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Their insights into biases in generative AI models are directly applicable to my research. I will use their findings to examine whether similar biases influence hiring algorithms. Specifically, I will reference their methods of identifying and quantifying biases in model outputs to analyze hiring tools. Similarly, Gichoya and coauthors (2023), published in The British Journal of Radiology, is another peer-reviewed source that explores biases in AI applications. I will use their discussion of bias in medical imaging AI to compare how algorithm biases manifest in different domains, using these similarities to inform my recommendations for bias mitigation in hiring algorithms.

Researchers have already explored how bias manifests in AI, acknowledging that the models used to train these systems often mirror the prejudices present in society (Duncan & Mcculloh, 2024). However, further research is needed to understand how these systems can be designed to actively prevent bias from occurring. A theoretical framework from STS, Actor-Network Theory

(ANT), will guide my analysis of how bias enters AI systems. ANT examines how social and technical networks interact to shape outcomes (Nickerson, 2024). I will use ANT to analyze how biases in AI hiring systems emerge from the interaction of technical, social, and organizational networks. Through case studies of biased AI outcomes in different industries, I will investigate the root causes of bias and the ethical implications of deploying these technologies. I will select case studies that highlight specific examples where AI systems produced biased decisions, such as hiring algorithms favoring male candidates. Additionally, I will consult experts in AI ethics and data governance, using their insights to propose methods for mitigating bias. These interviews will allow me to gather qualitative data on current challenges in developing fair AI systems. By focusing on the intersection of technology, society, and ethics, my research will contribute to ongoing discussions about the responsible development and deployment of AI, proposing actionable methods to mitigate bias and promote fairness in hiring.

**Conclusion**

My technical research aims to design a secure, scalable system that allows organizations to protect sensitive user data effectively. This design will provide practical solutions for strengthening data privacy and security in a way that adapts to evolving cyber threats and contributes to global cybersecurity resilience. Meanwhile, my STS research will deepen understanding of bias in AI hiring systems and present actionable strategies for mitigating these biases. Together, these projects address the challenges of security and fairness in data usage, contributing to both the technical and ethical areas of data management. By enhancing security frameworks and promoting fair AI practices, my research offers a pathway toward more trustworthy, responsible use of data. Solving these problems will promote a safer digital environment, protecting individuals' privacy, and ensuring equitable treatment in AI processes.

Works Cited

Alshammari, A., & Aldribi, A. (2021). Apply machine learning techniques to detect malicious network traffic in cloud computing. *Journal of Big Data*, *8*(1), 90. https://doi.org/10.1186/s40537-021-00475-1

Dastin, J. (2022). Amazon scraps secret AI recruiting tool that showed bias against women. In *Ethics of data and analytics* (pp. 296-299). Auerbach Publications.

Duncan, C., & Mcculloh, I. (2024). Unmasking bias in chat gpt responses. *Proceedings of the 2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 687–691. https://doi.org/10.1145/3625007.3627484

Huang, C., Zhang, Z., Mao, B., & Yao, X. (2023). An overview of artificial intelligence ethics. *IEEE Transactions on Artificial Intelligence*, *4*(4), 799–819. https://doi.org/10.1109/TAI.2022.3194503

Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2022). Trustworthy artificial intelligence: A review. *ACM Comput. Surv.*, *55*(2), 39:1-39:38. https://doi.org/10.1145/3491209

Gichoya, J. W., Thomas, K., Celi, L. A., Safdar, N., Banerjee, I., Banja, J. D., Seyyed-Kalantari, L., Trivedi, H., & Purkayastha, S. (2023). AI pitfalls and what not to do: Mitigating bias in AI. *The British Journal of Radiology*, *96*(1150), 20230023. https://doi.org/10.1259/bjr.20230023

Nadeem, A., Abedin, B., & Marjanovic, O. (2020). Gender bias in ai: A review of contributing factors and mitigating strategies. *ACIS 2020 Proceedings*. https://aisel.aisnet.org/acis2020/27

Nickerson, C. (2024, February 13). *Latour's Actor Network Theory*. https://www.simplypsychology.org/actor-network-theory.html

Parikh, R. B., Teeple, S., & Navathe, A. S. (2019). Addressing bias in artificial intelligence in health care. *JAMA*, *322*(24), 2377. https://doi.org/10.1001/jama.2019.18058

Roselli, D., Matthews, J., & Talagala, N. (2019). Managing bias in ai. *Companion Proceedings of The 2019 World Wide Web Conference*, 539–544. https://doi.org/10.1145/3308560.3317590

Sun, Y., Zhang, J., Xiong, Y., & Zhu, G. (2014). Data security and privacy in cloud computing. *International Journal of Distributed Sensor Networks*, *10*(7), 190903. https://doi.org/10.1155/2014/190903

Vasconcelos, M., Cardonha, C., & Gonçalves, B. (2018). Modeling epistemological principles for bias mitigation in ai systems: An illustration in hiring decisions. *Proceedings of the*

*2018 AAAI/ACM Conference on AI, Ethics, and Society*, 323–329. https://doi.org/10.1145/3278721.3278751