Amazing technologies—I caught the Computer Science bug at 17 when I took my first CS class in high school, and for the next 13 years, even as I majored in CS, I have remained awestruck by technologies. But I'm also concerned. The deepfake Obama looked *very* realistic. It only took a few seconds before I realized that it wasn't real, but that was already six years ago. Now, deepfakes have infiltrated almost every corner of our daily lives, and the technology has advanced to a frightening degree.

In my view, this has become a serious human rights issue that infringes on privacy and security. More concerning is that these incidents are likely to increase. The signs are evident: In 2023, around 500,000 video and voice deepfakes were shared globally on social media, a number projected to surge to 8 million by 2025 [1]. In 2024, in South Korea, deepfake sexual crimes targeted not only public figures but also ordinary individuals.

This urgent societal challenge motivates me to pursue research in trustworthy and responsible AI. But the vulnerabilities of AI are not limited to malicious uses like deepfakes. Even high-profile systems such as AlphaGo, in its 2016 match against Lee Sedol, have shown that when faced with unfamiliar scenarios, AI can produce irrational decisions. AlphaGo's repeated errors after the 78th move led to its defeat and highlighted how uncertainty and complexity can cause AI systems to fail unpredictably. I have observed similar issues firsthand, as certain models—due to limited adaptability in unstructured data environments—produced unexpected outcomes or exhibited biased predictions. Together, these examples—from malicious manipulation to unexpected failure—underscore that trustworthiness must be established at a fundamental level in AI design.

My research vision is to create responsible and trustworthy AI. To realize this vision, I identified four key elements in responsible and trustworthy AI design: security, robustness, interpretability, and fairness. These elements can be achieved by verifying data reliability, ensuring generalization beyond training distribution, developing transparent interpretation methods, and mitigating bias. Because AI is inherently data-centric, securely managing data and deriving meaningful insights are central to establishing trustworthiness. Thus, data security emerged as fundamental since it protects personal information and enables responsible data use. Accordingly, I concentrated on differential privacy (DP) as a primary technique.

My focus on DP began at the Center for Responsible AI at NYU, where I worked on DP-synthetic data generation techniques. DP preserves the statistical properties of data while safeguarding individual privacy, making it indispensable for trustworthy AI. However, early work did not sufficiently demonstrate DP's practical utility, leading to skepticism. To address this gap, I co-developed the *E pistemic Parity* metric, which evaluates whether research conclusions drawn from original data hold when replicated on DP-synthesized data. Moving beyond basic statistical comparisons and model performance metrics, our method provides rigorous evidence of DP's real-world relevance.

Using *Epistemic Parity*, I evaluated four DP-synthesizers—MST, AIM, PrivMRF, and PATECTGAN—on sensitive ICPSR social science data. I chose to replicate the Americans' Changing Lives study [2] and the Pierce and Quiroz [3] paper, which examines the impact of social support and conflict, and

replicated 100% of the original research findings under four privacy constraints. *Epistemic Parity* played a crucial role in quantitatively validating the reliability of DP-synthesized data and showed that meaningful research can be conducted using protected data. This empirical evidence establishes DP as a central methodology for data security and a critical step toward trustworthy AI systems,

However, data security alone is not sufficient. When exposed to out-of-distribution (OOD) scenarios or novel environments, models may produce unstable predictions and increased uncertainty, risking catastrophic outcomes in high-stakes domains. For instance, inaccurate predictions of new disease variants in the medicine pose direct threats to patient safety. Recognizing that robustness is essential for stable decision-making under uncertainty, I focused my subsequent research efforts on enhancing model robustness.

I compared a range of vision architectures—Convolution, Attention, and Hybrid—under identical training conditions and evaluated their performance on five ImageNet OOD datasets. Hybrid models achieved stronger robustness than Transformers and CNNs, though robustness varied with OOD scenarios. By highlighting where models falter, this work underscores the second mission—ensuring robustness—so that AI can make stable decisions, even in uncertainty.

Building on these insights, I refined my objective: to implement trustworthy AI by integrating robustness and security into a unified design framework. However, in complex and uncertain real-world settings, security and robustness alone do not suffice. In high-stakes domains such as healthcare, finance, and self-driving cars, model decisions directly affect human lives, making interpretability essential. Tools like LIME and SHAP help clarify simpler models, but more advanced techniques are needed for complex, context-dependent systems.

Fairness is the last principle—to abide by this principle, especially in CS, many questions need to be answered first and foremost, such as who to save first in case of an accident. Models trained on biased data may learn and perpetuate such bias. By applying reweighting algorithms to a loan repayment prediction model, I mitigated historical biases. Fairness ensures that AI respects societal values, making technology equitable, not exploitative.

We need to keep asking such challenging questions and seeking answers. That way, we can develop AI capable of handling very complex situations and uncertainties. In high-stakes domains, especially tasks requiring strategic and long-term approaches, answers to these questions are essential for using AI reliably and safely, just as the AI behind the wheel in a self-driving car needs to make very complex decisions. I also want to research ways for AI to produce optimal and ethical solutions in such complex and uncertain situations.

During my PhD, I will deepen these research directions by proposing and validating an AI design framework that unites security, robustness, fairness, and interpretability. Through this endeavor, I hope to establish new standards for trustworthy AI and ensure a responsible, positive influence on society.

References

[1] Ulmer, A., Tong, A., Ulmer, A., & Tong, A. (2023, May 31). Deepfaking it: America's 2024 election collides with AI boom. *Reuters*. https://www.reuters.com/world/us/deepfaking-it-americas-2024-election-collides-with-ai-boom-2023-05-30/

[2] House, James S. Americans' Changing Lives: Waves I, II, III, IV, and V, 1986, 1989, 1994, 2002, and 2011. Inter-university Consortium for Political and Social Research [distributor], 2018-08-22. https://doi.org/10.3886/ICPSR04690.v9

[3] Pierce, K. D., & Quiroz, C. S. (2019). Who matters most? Social support, social strain, and emotions. *Journal of Social and Personal Relationships, 36*(10), 3273-3292.