Amazing technologies—I caught the Computer Science bug at 17 when I took my first CS class. Over the next 13 years, even as I majored in CS, I've remained awestruck by technology's transformative power.

However, I'm also concerned. The Obama deepfake once seemed incredibly realistic, and while I realized it wasn't genuine after a few seconds, that was six years ago. Today, deepfakes infiltrate our daily lives, evolving to an alarming degree.

This phenomenon isn't merely a technological issue—it's a serious human rights issue that infringes on privacy and security, and incidents are likely to increase. In 2023, approximately 500,000 video and voice deepfakes were shared globally on social media, which may soar to 8 million by 2025 [1]. In 2024, deepfakes victimized both public figures and ordinary individuals in South Korea.

This urgent societal challenge motivates me to pursue research in trustworthy and responsible AI. Yet, malicious deepfakes aren't the only problem. Even well-graded systems can fail unpredictably. For example, AlphaGo famously faltered after the 78th move against Lee Sedol, revealing how unfamiliar scenarios push AI into irrational decisions. I've seen similar issues firsthand: certain models produce unexpected outcomes or exhibit biased predictions in unstructured data environments. These examples—from malicious manipulation to unexpected failures—underscore that trustworthiness must be central in AI design.

My research vision is to create responsible and trustworthy AI. To realize this, I identified four key elements: security, robustness, interpretability, and fairness. Achieving these requires verifying data reliability, ensuring out-of-distribution (OOD) generalization, developing transparent interpretation, and mitigating bias. Because AI is inherently data-centric, managing data and deriving meaningful insights are central to establishing trustworthiness. Thus, I focused on differential privacy (DP) as a foundational technique.

My focus on DP began at the Center for Responsible AI at NYU, where I worked on DP-synthetic data generation. DP preserves the statistical properties while protecting indivi dual privacy, making it indispensable for trustworthy AI. However, previous studies strug gled to demonstrate DP's practical utility, causing skepticism. To address this, I co-develo ped "Epistemic Parity", a metric evaluating whether conclusions drawn from original data hold when replicated with DP-synthesized data. Unlike basic comparisons, our method pro vides rigorous evidence of DP's real-world relevance.

Using Epistemic Parity, I evaluated four DP-synthesizers on sensitive ICPSR social science data. I reproduced the Americans' Changing Lives study [2] and the Pierce and Quiroz [3] paper, which examines the impact of social support and conflict, successfully replicating 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 demonstrated 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.

Still, 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 medicine pose direct threats to patient safety. Recognizing that robustness is essential for stable decision-making under uncertainty, I directed my subsequent research efforts to enhance 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 importance of robustness to ensure stable decision-making under uncertainty.

With robustness and security as anchors, I refined my objective: to implement trustworthy AI by integrating robustness and security into a unified design framework. However, in complex real-world settings, security and robustness alone are insufficient. In high-stakes domains such as healthcare, finance, and self-driving vehicles, model decisions directly affect human lives, making interpretability essential. Tools like LIME and SHAP help clarify simpler models, but more advanced techniques are required for complex, context-dependent systems.

Fairness forms the final pillar. To abide by this principle, particularly in CS, many foundational questions must be addressed first and foremost, such as determining priorities in life-and-death scenarios. Models trained on biased data may learn and perpetuate such bias. I've analyzed the COMPAS, a "risk assessment" tool, and mitigated bias in the loan repayment model by applying reweighting algorithms. Fairness ensures that AI respects societal values, making technology equitable, not exploitative.

We must continue posing such challenging questions and seeking answers. In doing so, we can develop AI capable of handling complexities 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 explore methods for AI to produce optimal and ethical solutions in similarly complex and uncertain conditions.

In sum, AI is extraordinary—arguably a "*brave new world*"—and I am eager to contribute to this unprecedented wave. The potential of AI is still largely untapped, and we've only witnessed a fraction of its capabilities. However, numerous questions must be answered before developing *truly* trustworthy AI systems. I'll play a part in answering them along with other scholars—this is why I must pursue a PhD.

My academic journey began with a youthful fascination for technology and has evolved into a focused commitment to creating AI that advances society responsibly. By refining privacy-preserving techniques, evaluating OOD robustness, and addressing fairness and interpretability, I've prepared to push the boundaries of responsible and trustworthy AI research. I'll build on this foundation, creating frameworks that ensure AI's benefits are shared ethically. In doing so, I'll seek to participate in the next wave of AI innovation and guide it toward responsible outcomes.

References

[1] Ulmer, A., Tong, A., Ulmer, A., & Tong, A. (2023, May 31). Deepfaking It: Americ a's 2024 Election Collides with AI Boom. Reuters. https://www.reuters.com/world/us/deepf aking-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 [dis tributor], 2018-08-22. https://doi.org/10.3886/ICPSR04690.v9

[3] Pierce, K. D., & Quiroz, C. S. (2019). Who Matters Most? Social Support, Social St rain, and Emotions. Journal of Social and Personal Relationships, 36(10), 3273-3292.