Research Statement

Xi Li

Department of Computer Science and Engineering, The Pennsylvania State University

Email: XiLi@psu.edu | *Phone:* (814) 777-6667 | *Website:* lixi1994.github.io

1 Introduction

Trustworthy Artificial Intelligence (AI) has become a significant concern as become increasingly integral to various applications—ranging from personal diagnosis and drug discovery to autonomous vehicles and chatbots. Beyond the efficiency and effectiveness of ML models, their ethical dimensions, including robustness, privacy, and fairness, demand equal attention. These models must be resilient to errors and adversarial inputs, safeguard user's sensitive information, and ensure equitable treatment across diverse user groups. Consequently, as demonstrated in Fig 1, **my research vision is to bridge AI with interdisciplinary research by developing trustworthy and reliable AI systems that not only advance technology but also serve the greater social good.**

My **previous research** focuses on the **robustness** of machine learning, with specific studies on the data poisoning (DP) attacks on ML systems and the defenses against such attacks. This body of work, illustrated in Fig. 2, follows the adversarial evolution between attackers and defenders alongside the advancement of ML technology. My journey began with defending against label-flipping DP attacks, I worked on extending, decoding, and mitigating stealthy backdoor poisoning attacks against DNNs, and studied adversarial threats empowered by Foundation models (FMs) against classic ML frameworks. These works addressed gaps left by previous research, explored DNN robustness in novel domains, and offered novel insights in adversarial attacks.

With the advent of Foundation Models (FMs) like ChatGPT and diffusion models, the learning capabilities and complexity of ML have broadened, finding applications across numerous domains. This expansion amplifies the responsibility concerns associated with ML models. In my **future research**, I aim to (1) enhance the **robustness** of traditional ML systems integrated with FMs and the FMs themselves; (2) assess and address the broader **ethical** concerns surrounding ML models, such as **privacy**, **fairness**, and **interpretability**; (3) explore the **interdisciplinary** application of responsible AI, ensuring transparent, interpretable, and equitable decision-making processes.

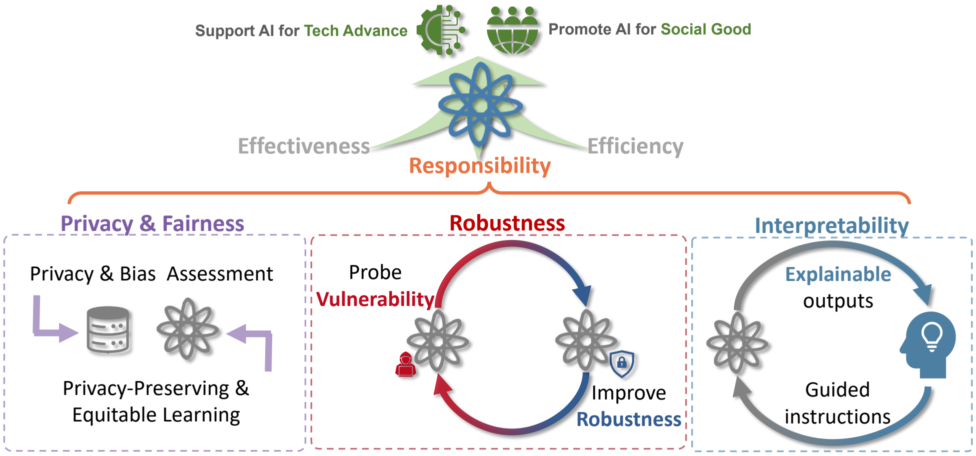

Figure 1: My future research direction in the development of trustworthy AI.

2 Past Research

My past research, as shown in Fig. 2, is consistent with the evolution of adversariality between attackers and defenders – devising defense-invisible attacks on ML models and defenses against emerging attacks. It also aligns with the development of machine learning – from traditional deep learning to foundation models.

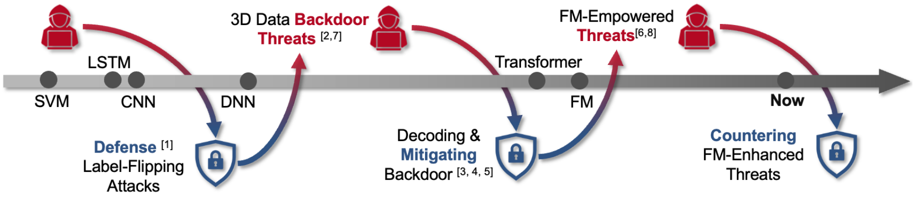
3D Data Backdoor Threats [2,7]
FM-Empowered Threats[6,8]
LSTM
Transformer
SVM
CNN
DNN
FM
Now
Defense [1] Label-Flipping Attacks
Decoding & Mitigating Backdoor [3, 4, 5]
Countering FM-Enhanced Threats

Figure 2: My research journey in adversarial machine learning.

I began the journey with an unsupervised defense against label-flipping DP attacks [1], notably effective in spam and fraud detection, aimed at degrading overall performance of the detector. We addressed the practical DP scenario where, if DP is present, the poisoned samples are an a priori unknown subset of the training set without a clean validation set for the defender. We formulated the anomaly detection problem as a Bayesian Information Criterion (BIC) minimization process. The intuition behind this is that reassigning and representing the likely poisoned samples according to the density models of other classes would increase the total data likelihood, and removing or revising the components representing these samples would decrease the total model cost, aligning with the goal of minimizing BIC. Additionally, our detection strategy is unsupervised: its performance does not rely on the choice of hyperparameters.

Then, I progressed to extending, decoding, and detecting/mitigating backdoor poisoning attacks against DNNs. Backdoor attacks aim for targeted misclassifications on specific inputs without compromising the overall model performance, rendering the attack stealthy. Pioneering in exploring DNN vulnerabilities to backdoor threats in 3D data, such as point cloud classification[9] and video recognition[3], we faced unique challenges due to the differing structures and extraction strategies of 3D data. Our solutions – embedding a point cluster in 3D PCs and crafting temporal triggers in video data – set the stage for developing sophisticated defense mechanisms for DNN robustness in new domains. Moreover, while existing backdoor research primarily emphasizes stealth and adaptability, we explored the fundamental characteristics of backdoor attacks, discovering unique neuron activation patterns triggered by backdoors. This insight led to an unsupervised technique for real-time detection of triggered instances. Further, we were the first to identify and analyze the distribution alteration property of backdoor attacks, showing how triggers change internal activation distributions. Based on this, we theoretically validated the monotonicity of classification accuracy concerning distribution divergence and introduced a novel post-training mitigation strategy that does not require altering model parameters[8], proving particularly effective with limited clean data.

In the era of FMs, we made a preliminary exploration on the adversarial threats empowered by the FMs against classic ML frameworks [4, 5, 6] – backdoors transferred from the FMs to the downstream models follow distinct attack patterns than traditional poisoning-based backdoor attack, thus are able to evade the existing backdoor defenses.

3 Future Research

Recently, the extensive integration of FMs like ChatGPT into real-world applications and classic ML frameworks has introduced potential risks to traditional ML models and raised security concerns about the FMs themselves. Moreover, the expansion of large models amplifies the responsibility concerns associated with ML models — we aim for AI to adhere to human standards. Therefore, my **short-term** research will focus on enhancing the robustness of FM-integrated ML systems and FMs themselves. My long-term goal is to broaden my research scope to include other ethical aspects such as privacy and fairness, developing and deploying trustworthy AI across interdisciplinary domains to ensure transparent, interpretable, and equitable decision-making processes. My **research goal** is to support emerging technologies and foster a human-centric, ethical AI ecosystem, expanding research into the societal impacts and ethics of AI.

3.1 Delving into ML robustness: FMs are enhancing traditional ML systems by addressing their limitations and improving performance through synthetic data generation, knowledge distillation, and feature extraction, among other techniques. Yet, this integration raises new robustness challenges, as highlighted in [4, 5]. My short-term research will focus on: (1)Enhancing the robustness of FM-integrated ML systems against emerging threats. (2)Investigating the resilience of FMs in both single-agent and multi-agent scenarios. Building on my expertise in DNN vulnerabilities and robustness [3, 9, 1, 10, 7, 8], and my knowledge of widely-used FMs [4, 5, 6], I plan to adapt my approach of examining adversarial dynamics to the FM era. Traditional defense methods, often reliant on gradient computation and fine-tuning, face scalability issues with FMs. Hence, I aim to devise novel defense strategies leveraging the unique reasoning abilities of LLMs to identify and counteract manipulations leading to abnormal outputs. The current deployment of FMs typically involves a single-agent mode, where tasks are handled independently. However, the rise of multi-agent systems, where multiple FMs collaborate, introduces complex interactions that could potentially uncover and exploit new vulnerabilities, marking a departure from traditional adversarial dynamics.

3.2 Broadening to ML responsibility: The widespread adoption of FMs raises significant responsibility concerns about privacy and fairness. For instance, LLMs like ChatGPT risk leaking sensitive data and perpetuating biases from their internet-sourced training materials. Additionally, the complexity of FMs complicates the transparency of their decision-making processes, making it more challenging to address these issues. Moving forward, my **long-term** will tackle privacy and fairness issues in FM-integrated ML systems, including assessing privacy risks during model deployment and developing algorithms for data privacy protection. I will also focus on creating unbiased datasets and ensuring fair and equitable outcomes in collaborative AI scenarios. Additionally, I plan to work on improving the interpretability and transparency of AI, making the decision-making processes of AI more understandable to a broader range of stakeholders. This comprehensive strategy aims to address ethical dilemmas and foster responsible, inclusive AI across various domains.

3.3 Interdisciplinary Applications of trustworthy AI: Trustworthy AI, known for its reliability, ethical design, and transparency, emerges as a key solution to address social challenges. Therefore, building on my preliminary exploration [2], my future research aims to create AI systems that promote social good and technological advancement. For instance, I could partner with: 1. **Biomedical** scientists to develop reliable diagnostic tools and personalized medicine, thereby improving patient care, ensuring equitable treatment, and making decisions more interpretable. 2. **IoT** professionals to enhance user privacy and defend against cyber threats. 3. **Finance** data scientists to create anti-fraud algorithms, contributing to secure and transparent financial transactions. 4. Engineers in the **autonomous driving** field to ensure the reliability and safety of vehicular AI. 5. **Advanced Robotics** experts to improve AI for greater fault tolerance in high-risk areas like surgery and space exploration.

References

[1] Xi Li, David J. Miller, Zhen Xiang, and George Kesidis. A BIC based mixture model defense against data poisoning attacks on classifiers. *IEEE TKDE*, 2024.

[2] Xi Li and Jiaqi Wang. Position paper: Assessing robustness, privacy, and fairness in federated learning integrated with foundation models. under review, 2024.

[3] Xi Li, Songhe Wang, Ruiquan Huang, Mahanth Gowda, and George Kesidis. Temporal-distributed backdoor attack against video based action recognition. *AAAI*, 2024.

[4] Xi Li, Songhe Wang, Chen Wu, Hao Zhou, and Jiaqi Wang. Backdoor threats from compromised foundation models to federated learning. *FL@FM with NeurIPS,*, 2023.

[5] Xi Li, Chen Wu, and Jiaqi Wang. Unveiling backdoor risks brought by foundation models in heterogeneous federated learning. 2024. PAKDD.

[6] Xi Li, Chen Wu, and Jiaqi Wang. Vulnerabilities of foundation model integrated federated learning systems under adversarial threats. under review, 2024.

[7] Xi Li, Zhen Xiang, David J. Miller, and George Kesidis. Test-time detection of backdoor triggers for poisoned deep neural networks. In *IEEE ICASSP*, 2022.

[8] Xi Li, Zhen Xiang, David J. Miller, and George Kesidis. Backdoor mitigation by correcting the distribution of neural activations. *under review of Neurocomputing*, 2023.

[9] Zhen Xiang, David J. Miller, Siheng Chen, Xi Li, and George Kesidis. A Backdoor Attack against 3D Point Cloud Classifiers. *ICCV*, 2021.

[10] Zhen Xiang, David J. Miller, Siheng Chen, Xi Li, and George Kesidis. Detecting backdoor attacks against point cloud classifiers. In *ICASSP*, 2022.