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 deep neural networks (DNNs) become integral to safety-critical systems like vehicle autopilots and medical diagnostics. Such systems require a high degree of safety and robustness, and failure in these systems could result in significant property damage or even loss of life. On the other hand, the robustness of machine learning (ML) models attracts attention as a result of the prevalence of adversarial attacks. The complexity and potent learning capabilities of DNNs can lead to errors when exposed to manipulated inputs or compromised training data, such as an autopilot system mistaking a stop sign for a speed limit sign due to subtle adversarial perturbations. Further, foundation models (FMs) like ChatGPT and Dall-E have expanded ML's learning capabilities and complexity, and thus are more vulnerable to adversarial attacks. Despite their widespread use, research into the robustness of FMs is still limited. The risks are tangible: manipulated prompts could induce ChatGPT to output harmful content. In a word, **robustness is not an optional attribute but a critical feature** that ensures machine learning models can be trusted in real-world applications.

Hence, my **research vision** is centered on integrating AI research with research in other areas, emphasizing its influence on both technology and society. My **research goal** is to develop trustworthy and reliable AI systems to support AI for *technology advancement* and promote AI for *social good*. Motivated by the critical need for robust AI in real-world applications, **my research focuses on adversarial machine learning**, with specific studies on the data poisoning (DP) attacks on ML systems and the defenses against such attacks. My research, as shown in Fig. 1, follows the evolution of adversariality between attackers and defenders and the development of ML. Starting from 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 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.

In the era of FMs, security concerns of FMs attract attentions, as well as the robustness of traditional DNNs. In my **future research**, I will focus on (1) enhancing the robustness of traditional DNNs against unsolved threats and emerging threats empowered by FMs; (2) exploring the robustness of FMs, following the evolutionary pattern of adversariality observed in traditional DNNs; (3) contribute to next generation of adversarial learning by automatically promoting the adversariality between attackers and defenders.

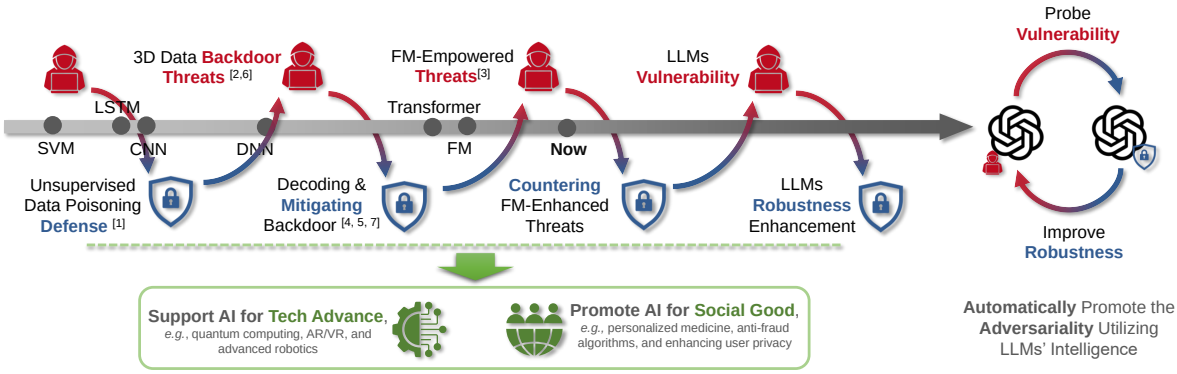

Figure 1: Overview of My Research: Past Achievements and Future Directions From Adversarial Learning to Supporting AI for Technology Advancement and Promoting AI for Social Good.

2 Past Research

My research on secure machine learning is twofold: (1) probing vulnerability of ML models under potential threats; (2) enhancing robustness of ML models against adversarial attacks. My research pattern, as shown in Fig. 1, 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. Specifically, I started the journey with defense against label-flipping DP attacks and progressed to extending, decoding, and detecting/mitigating stealthy backdoor poisoning attacks against DNNs. In the era of FMs, my preliminary research has focused on assessing the risks associated with integrating FMs into classic ML frameworks.

2.1 Unsupervised Defense for Label-Flipping Attacks: Classic poisoning attacks aim to degrade the overall performance of the victim model by inserting malicious samples into the training set. For example, label-flipping poisoning attacks are effective attacks against spam detection, fraud detection, and sentimental analysis, where the labels of a training subset are maliciously modified. A practical and challenging attacking scenario, named embedded DP attack, remains largely unsolved. In the scenario, if DP is present, the poisoned samples are an a priori unknown subset of the training set, and with no clean validation set available for the defender. We propose the first unsupervised defense strategy to solve it, based on Bayesian Information Criterion (BIC) and mixture models [2]. Our method relies on two key insights: First, we propose that poisoned samples create distinct sub-groups within their labeled classes, and by using mixture modeling, we can represent the dataset accurately to potentially isolate these samples. Second, we expect the dataset's overall likelihood to increase if poisoned samples are correctly reclassified. We then either remove or adjust these poisoned components, introducing changes in model complexity. As the BIC balances data fit and model complexity, we thus aim to identify poisoned samples and components by enhancing data fit while reducing complexity, in line with minimizing BIC. Our unsupervised detection strategy offers significant advantages over current DP detection and mitigation methods: 1) its performance does not rely on the choice of hyper-parameters; 2) is empirically proven effective under strong attacks.

2.2 Stealthy Backdoor Poisoning Threats in 3D Data: Classic DP attacks, while effective, lack stealth due to noticeable validation accuracy drops during training. Backdoor attacks, a novel DP attack originally proposed against 2D image data, attracts attention. It aims for *targeted* misclassifications on specific inputs without compromising overall model performance, rendering the attack in a stealthy way. As AI rapidly evolves, DNN applications have expanded into 3D data areas like autonomous vehicles, industrial robotics, and augmented reality/virtual reality. Consequently, our research shifts to these emerging domains, being the first to investigate DNN vulnerabilities to sophisticated backdoor threats in 3D point cloud (PC) classification [8] and video recognition systems [3]. Devising backdoor triggers for 3D PC presents challenges: 1) nonexistence of "pixels"; 2) difference in feature extraction; 3) resistance to pre-processing and remaining scene-realistic. Our solution is to embed a point cluster as a backdoor into the PC, which is optimized for spatial location and local geometry to be captured by PC classifiers and survive through pre-processing and anomaly detectors. On the other hand, while videos are image sequences, direct backdoor extensions can not evade existing backdoor defenses. Hence, we leverage the temporal dimension for stealthier attacks, crafting triggers that are imperceptible and distributed over time. Current defenses, limited to frame-by-frame analysis, miss these temporal patterns. Our attack explores time-variant transforms like Fourier transform and Wavelet transform. Our research also uncovered "collateral damage," where certain perturbations could unintentionally trigger the attack, which is undiscovered by existing frequency-domain attacks. Our results encourage the development of advanced defense strategies to ensure the robustness of DNNs in innovative domains.

2.3 Decoding and Mitigating Backdoor Attacks: While most backdoor research focuses on attack stealth, flexibility, adaptation, or defense, there is limited exploration of the basic properties of backdoor attacks. Our preliminary study [6] reveals that backdoor triggers activate distinct neuron patterns in DNNs. Leveraging this, we propose an unsupervised technique for detecting triggered instances at test time, filling gaps left by existing detectors. Then we delve deeper into the properties of backdoor attacks. We are the first to investigate

the *distribution alteration property* of backdoor attacks – the learned backdoor trigger causes a change in the distribution of internal activations for test instances with the trigger, compared to that for backdoor-free instances. Then we *theoretically* prove the *monotonicity* of classification accuracy on backdoor-trigger instances with distribution divergence in DNN internal layer feature space, and propose a post-training backdoor attack mitigation approach by correcting distribution of neural activation using reverse-engineered triggers, *without* modifying trainable model parameters [7]. In post-training scenarios, without the original data, defenders rely on a compromised model and minimal clean data. Our method outperforms other methods, especially when the defender possesses limited clean data, since our method do not involve parameter tuning.

2.4 Adversarial Threats Empowered by Foundation Models: FMs lead to a new paradigm in machine learning. The vast parameters of FMs enhances their learning capacities, outperforming traditional DNNs across diverse domains, e.g., natural language processing. However, their extraordinary complexity also increases their vulnerability to adversarial attacks. For example, backdoor attackers traditionally use poisoned training to guide the victim DNN learn a mapping from a specific trigger to the attacker-chosen target class. However, LLMs can learn these mappings through in-context learning during inference, without poisoned training. Hence, there is a transformation in the approaches to implementing attacks and defenses. We made a preliminary exploration on the transformation in [4, 5]. In these works, we demonstrate that, for ML frameworks integrated with FMs, backdoors transferred from the FMs to the downstream models are more effective than traditional poisoning-based backdoor attacks. Specifically, FMs enhance federated learning (FL) with pre-training on synthetic data. If compromised, these FMs implant backdoors into the initial model, which, after fine-tuning with client data during FL, quickly converges and retains the backdoor, due to the effective initial setup. This method outperforms classic FL backdoor attacks, particularly in large-client scenarios, and stealthily bypasses current federated 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 regarding FMs themselves. My **short-term** research will therefore focus on enhancing the robustness of both traditional DNNs and FMs. The **long-term** goal is to develop a novel adversarial learning strategy where the evolution of adversariality is automatically promoted utilizing the LLMs' intelligence, thus achieving self-improving robustness. Based on my expertise in adversarial machine learning, my **research vision** is supporting emerging technologies and creating a human-centric, ethical AI ecosystem, expanding research into the societal impacts and ethics of AI.

3.1 Delving into AI Robustness and Advancing Next Generation of Adversarial Learning: Despite that FMs enhance ML capabilities, traditional DNNs are crucial in scenarios with limited resources, real-time demands, and privacy concerns. Therefore, the robustness of both traditional DNNs and FMs is garnering attention. Apart from existing threats, new challenges are emerging, enhanced by FMs as demonstrated in [4, 5]. My **short-term** research involves: (a) Improving **robustness** of traditional **DNNs** against existing threats and emerging threats; (b) Enhancing the **robustness** of **FMs** themselves. To address (a), I will focus on backdoor attacks and address the unresolved question of basic backdoor attack properties: "Do poisoned DNNs exhibit a common pattern with both universal and sample-specific backdoor triggers?" We have explored a unique property of backdoor attacks using universal triggers in [7]. Building on this, I plan to devise an efficient defense strategy applicable to all backdoor attack types, leveraging the identified common attack pattern. This strategy aims to fill the gaps left by prior research [1]. Besides, I plan to adapt my proven defense methods [2, 9, 6, 7] to solve the FM-enhanced threats. For solving (b), following my previous research pattern, I will focus on LLMs and thoroughly investigate their susceptibility to backdoor attacks on generative tasks, a relatively unexplored area, setting a benchmark for LLM robustness. This work is built on my background in studying DNN vulnerabilities [3, 8] and familiarity with widely-used FMs [4, 5]. However, the transition to FMs presents a critical challenge for defending them against attacks – traditional defenses, usually involve gradient computation and fine-tuning, don't scale well with LLMs. In light of these challenges, I propose to explore novel defense mechanisms using unique capabilities of LLMs, such as "chain of thought", which could help detect the unreasonable mapping from manipulated inputs to

abnormal outputs. My previous research sheds light on the direction, where I explored the basic properties of backdoor attacks from a unique perspective, and proposed an effective defense strategy without tuning model parameters [6, 7].

In current applications, LLMs are usually deployed in a "single-agent" mode, handling tasks individually. However, the emerging "multi-agent" trend has created ecosystems where multiple LLMs collaborate to solve a task. This inspires my **long-term** goal of **developing evolutionary adversariality in LLMs**. As shown in Fig. 1, my previous research on ML robustness evolves with the adversarial dynamics between attackers and defenders, which is promoted by **manual intervention**. I aim to **automate this evolution** by deploying LLMs as both attackers and defenders in a "multi-agent" setup, as shown in Fig. 1. Their iterative interactions will mutually enhance their ability to detect and exploit vulnerabilities, leading to a self-improving AI security system.

3.2 Supporting AI for Technology Advancement: As AI becomes integral to critical tech areas, there's a need for systems that are not only advanced but also reliable and safe. My work will focus on enhancing AI robustness in these fields to create public trust and facilitate their integration into both daily and high-stakes environments, including: (1) **Quantum Computing**: Crafting robust AI algorithms for more reliable quantum computations. (2) **Advanced Robotics**: Improving AI for greater fault tolerance and accuracy in high-risk areas like surgery and space exploration. (3) **AR/VR**: Bolstering AI resilience in AR/VR for consistent and reliable user experiences.

3.3 Promoting AI for Social Good: Facing social challenges like user privacy and financial integrity, trustworthy AI, known for its reliability, ethical design, and transparency, emerges as a key solution. Therefore, my future research also aims to promote AI for social good, creating trustworthy AI systems that address diverse societal challenges. For instance, I could partner with (1) **biomedical** scientists on reliable diagnostic tools and personalized medicine, improving patient care; (2) **IoT** professionals on enhancing user privacy and defending cyber threats; (3) **finance** data scientists on developing anti-fraud algorithms, contributing to the security of financial transactions; (4) engineers in the **auto-driving** field to ensure vehicular AI reliability and safety.

3.4 Interdisciplinary Research: As aforementioned, I could promote AI for technology advancement and social good through interdisciplinary collaboration in fields such as quantum computing, robotics, AR/VR, biomedical, and finance. Finally, I can cooperate with **computer science** fellow researchers for a deeper dive into advanced adversarial strategies, bolstering the robustness of AI systems against emerging threats.

References

[1] Xi Li, David J. Miller, and George Kesidis. A general framework of reverse-engineering backdoor trigger in the embedded feature space. Manuscript in preparation, 2023.

[2] Xi Li, David J. Miller, Zhen Xiang, and George Kesidis. A BIC based mixture model defense against data poisoning attacks on classifiers. *IEEE MLSP*, 2023. A complete version of the paper is under the second round review of IEEE TKDE.

[3] Xi Li, Songhe Wang, Ruiquan Huang, Mahanth Gowda, and George Kesidis. Temporal-distributed backdoor attack against video based action recognition. *Under review of AAAI*, abs/2308.11070, 2023.

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

[5] Xi Li, Chen Wu, and Jiaqi Wang. Unveiling backdoor risks brought by foundation models in heterogeneous federated learning. Under review of PAKDD, 2023.

[6] 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.

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

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

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