Siyuan Cheng

cheng535@purdue.edu Homepage Google Scholar LinkedIn

RESEARCH INTEREST

- My research expertise lies in the realm of trustworthy machine learning, with a specific focus on adversarial/backdoor attacks and defenses, across various domains, including computer vision, natural language processing, self-supervised learning, and federated learning.
- My current focus is on real-world applications, e.g., large language models (LLMs) and diffusion models. I am actively engaged in exploring and addressing the intricate security and privacy concerns in these sophisticated systems.

EDUCATION

8/2021 – Present **Purdue University,** West Lafayette, IN, USA
Ph.D. Student, Computer Science

- Advisor: Prof. *Xiangyu Zhang*

9/2016 – 7/2020 **Shanghai Jiao Tong University (SJTU),** Shanghai, China
Bachelor, Computer Science and Technology (Artificial Intelligence)

- *IEEE Honor Class*

PUBLICATIONS

*Note: * denotes equal contribution*

S&P 2024 ***ODSCAN*: Backdoor Scanning for Object Detection Models**
Siyuan Cheng*, Guangyu Shen*, Guanhong Tao, Kaiyuan Zhang, Zhuo Zhang, Shengwei An, Xiangzhe Xu, Yingqi Liu, Shiqing Ma, Xiangyu Zhang

S&P 2024 **Exploring the Orthogonality and Linearity of Backdoor Attacks**
Kaiyuan Zhang*, **Siyuan Cheng***, Guangyu Shen, Guanhong Tao, Shengwei An, Anuran Makur, Shiqing Ma, Xiangyu Zhang

CVPR 2024 ***LOTUS*: Evasive and Resilient Backdoor Attacks through Sub-Partitioning**
Siyuan Cheng, Guanhong Tao, Yingqi Liu, Guangyu Shen, Shengwei An, Shiwei Feng, Xiangzhe Xu, Kaiyuan Zhang, Shiqing Ma, Xiangyu Zhang

NDSS 2023 ***BEAGLE*: Forensics of Deep Learning Backdoor Attack for Better Defense**
Siyuan Cheng, Guanhong Tao, Yingqi Liu, Shengwei An, Xiangzhe Xu, Shiwei Feng, Guangyu Shen, Kaiyuan Zhang, Qiuling Xu, Shiqing Ma, Xiangyu Zhang

NeurIPS 2023 ***Django*: Detecting Trojans in Object Detection Models via Gaussian Focus Calibration**
Guangyu Shen*, **Siyuan Cheng***, Guanhong Tao, Kaiyuan Zhang, Yingqi Liu, Shengwei An, Shiqing Ma, Xiangyu Zhang

AAAI 2021 **Deep Feature Space Trojan Attack of Neural Networks by Controlled Detoxification**
Siyuan Cheng, Yingqi Liu, Shiqing Ma, Xiangyu Zhang

S&P 2024 **On Large Language Models' Resilience to Coercive Interrogation**
Zhuo Zhang, Guangyu Shen, Guanhong Tao, **Siyuan Cheng**, Xiangyu Zhang

AAAI 2024 **_Elijah_: Eliminating Backdoors Injected in Diffusion Models via Distribution Shift**
Shengwei An, Sheng-Yen Chou, Kaiyuan Zhang, Qiuling Xu, Guanhong Tao, Guangyu Shen, **Siyuan Cheng**, Shiqing Ma, Pin-Yu Chen, Tsung-Yi Ho, Xiangyu Zhang

Security 2023 **Hard-label Black-box Universal Adversarial Patch Attack**
Guanhong Tao, Shengwei An, **Siyuan Cheng**, Guangyu Shen, Xiangyu Zhang

S&P 2023 **_ImU_: Physical Impersonating Attack for Face Recognition System with Natural Style Changes**
Shengwei An, Yuan Yao, Qiuling Xu, Shiqing Ma, Guanhong Tao, **Siyuan Cheng**, Kaiyuan Zhang, Yingqi Liu, Guangyu Shen, Ian Kelk, Xiangyu Zhang

ICLR 2023 **_FLIP_: A Provable Defense Framework for Backdoor Mitigation in Federated Learning**
Kaiyuan Zhang, Guanhong Tao, Qiuling Xu, **Siyuan Cheng**, Shengwei An, Yingqi Liu, Shiwei Feng, Guangyu Shen, Pin-Yu Chen, Shiqing Ma, Xiangyu Zhang
ECCV 2022 AROW Workshop 🏆 _Best Paper Award_

CVPR 2023 **Detecting Backdoors in Pre-trained Encoders**
Shiwei Feng, Guanhong Tao, **Siyuan Cheng**, Guangyu Shen, Xiangzhe Xu, Yingqi Liu, Kaiyuan Zhang, Shiqing Ma, Xiangyu Zhang

CVPR 2023 **_MEDIC_: Remove Model Backdoors via Importance Driven Cloning**
Qiuling Xu, Guanhong Tao, Jean Honorio, Yingqi Liu, Shengwei An, Guangyu Shen, **Siyuan Cheng**, Xiangyu Zhang

NeurIPS 2023 BUGS Workshop **D^3: Detoxing Deep Learning Dataset**
Lu Yan, **Siyuan Cheng**, Guangyu Shen, Guanhong Tao, Kaiyuan Zhang, Yunshu Mao, Xiangyu Zhang

FSE 2023 **_PEM_: Representing Binary Program Semantics for Similarity Analysis via A Probabilistic Execution Model**
Xiangzhe Xu*, Zhou Xuan*, Shiwei Feng, **Siyuan Cheng**, Yapeng Ye, Qingkai Shi, Guanhong Tao, Le Yu, Zhuo Zhang, Xiangyu Zhang

ISSTA 2023 **Improving Binary Code Similarity Transformer Models by Semantics-driven Instruction Deemphasis**
Xiangzhe Xu, Shiwei Feng, Yapeng Ye, Guangyu Shen, Zian Su, **Siyuan Cheng**, Guanhong Tao, Qingkai Shi, Zhuo Zhang, Xiangyu Zhang

AAAI 2021 **Towards Feature Space Adversarial Attack by Style Perturbation**
Qiuling Xu, Guanhong Tao, **Siyuan Cheng**, Xiangyu Zhang

ICML 2021 **Backdoor scanning for deep neural networks through k-arm optimizationn**
Guangyu Shen*, Yingqi Liu*, Guanhong Tao, Shengwei An, Qiuling Xu, **Siyuan Cheng**, Shiqing Ma, Xiangyu Zhang

AsiaCCS 2020 **Rational Manager in Bitcoin Mining Pool: Dynamic Strategies to Gain Extra Rewards**
Feifan Yu, Na Ruan, **Siyuan Cheng**

PRE-PRINTS

Note: *denotes equal contribution

02/2024 **Rapid Optimization for Jailbreaking LLMs via Subconscious Exploitation and Echopraxia**
Guangyu Shen*, **Siyuan Cheng***, Kaiyuan Zhang, Guanhong Tao, Shengwei An, Lu Yan, Zhuo Zhang, Shiqing Ma, Xiangyu Zhang

01/2024 **Opening A Pandora's Box: Things You Should Know in the Era of Custom GPTs**
Guanhong Tao*, **Siyuan Cheng***, Zhuo Zhang, Junmin Zhu, Guangyu Shen, Xiangyu Zhang

11/2022 **Backdoor vulnerabilities in normally trained deep learning models**
Guanhong Tao, Zhenting Wang, **Siyuan Cheng**, Shiqing Ma, Shengwei An, Yingqi Liu, Guangyu Shen, Zhuo Zhang, Yunshu Mao, Xiangyu Zhang

6/2022 **Deck: Model hardening for defending pervasive backdoors**
Guanhong Tao, Yingqi Liu, **Siyuan Cheng**, Shengwei An, Zhuo Zhang, Qiuling Xu, Guangyu Shen, Xiangyu Zhang

EXPERIENCE

8/2021 – Present **Purdue University,** West Lafayette, IN, USA
Research Assistant

- Advisor: Prof. *Xiangyu Zhang*

10/2020 – 4/2021 **Hitachi Shanghai Trading Co Ltd,** Shanghai, China
Research Intern

7/2019 – 9/2019 **Purdue University,** West Lafayette, IN, USA
Research Intern

- Advisor: Prof. *Xiangyu Zhang*

INVITED TALKS

12/2023 **Enhance Trigger Inversion for Better Defense Against Backdoor Attacks**
Guest lecture at University of Massachusetts, Amherst

- COMPSCI 360: Introduction to Computer and Network Security
- Invited by Prof. Shiqing Ma

3/2023 ***BEAGLE*: Forensics of Deep Learning Backdoor Attack for Better Defense**
NDSS 2023, San Diego, CA, USA

- Video: YouTube

2/2021 **Deep Feature Space Trojan Attack of Neural Networks by Controlled Detoxification**
AAAI 2021, Virtually

- Video: Slideslive

SERVICES

Conference Reviewer

- IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR**): 2024
- Annual Meeting of the Association for Computational Linguistics (**ACL**): 2024

Sub-reviewer

- ACM Conference on Computer and Communications Security (**CCS**): 2021, 23, 24
- USENIX Security Symposium: 2022
- International Conference on Software Engineering (**ICSE**): 2023
- International Conference on Automated Software Engineering (**ASE**): 2023