Feng Wei

fengfeng.wf@gmail.com https://akafengfeng.github.io/

Research Interests

Feng Wei is an AI Security and Safety Researcher at the Artificial Intelligence Institute, China Academy of Information and Communications Technology (CAICT), Beijing. He received his Ph.D. in Computer Science and Engineering from the University at Buffalo, The State University of New York (SUNY) in 2025, where his research was supported by the Amazon Research Award and published at top AI and security conferences.

His research focuses on AI safety benchmarking, Agetitc AI security, LLM code security, cybersecurity evaluation of frontier models, and trustworthy machine learning. He also contributes to AI safety governance and standardization policy research for the National Development and Reform Commission (NDRC), the Ministry of Industry and Information Technology (MIIT), and the National Data Administration (NDA).

Experience

AI Security and Safety Researcher **May 2025 – Present**
Artificial Intelligence Institute, CAICT *Beijing, China*

- AI safety benchmarking for frontier models, evaluating cybersecurity capabilities and CBRN risks.
- Agentic AI security, focusing on skill uncertainty, memory evolution, and GUI/CLI interfaces.
- AI safety governance and standardization policy for the National Development and Reform Commission (NDRC), the Ministry of Industry and Information Technology (MIIT), and the National Data Administration (NDA).

Graduate Research Assistant **June 2021 – February 2025**
University at Buffalo, SUNY *New York, USA*

- Advisor: Prof. Hongxin Hu.
- Designed explainable deep learning methods for network intrusion detection (xNIDS), outperforming baseline explanation methods (LIME, SHAP, LRP, Integrated Gradients) in fidelity, sparsity, completeness, and stability.
- Developed defense rule generation methodologies for active intrusion response compatible with heterogeneous defense tools.
- Proposed robust training and data augmentation strategies (GAN-based) for DL-NIDS to withstand distribution shifts and adversarial attacks.
- Designed and trained security foundation models for NIDS integrated with LLMs for explainable and robust attack detection.
- Project received the Amazon Research Award (ARA) for AI and Information Security.

AI and Security Research Intern **May 2023 – August 2023**
Mitsubishi Electric Research Laboratories (MERL) *Cambridge, MA, USA*

- Collaborated with MERL researchers on developing robust AI methods for cybersecurity applications.

Education

Ph.D. in Computer Science and Engineering **University at Buffalo, SUNY**
Advisor: Prof. Hongxin Hu *June 2018 – February 2025*

B.E. in Automation **Xi'an Jiaotong University, 2014**

Selected Papers

- Yusheng Zhao, Jian Zhao, Tianle Zhang, **Feng Wei**, and Xuelong Li. "The Mark Fades: Adaptive Evolutionary Paraphrase-based Attack against LLM Watermarks." In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026.
- Lixun Ma, Ruolong Ma, Bei Wang, **Feng Wei**, Zhenguang Liu, Lorenzo Cavallaro, and Wentao Chen. "Rethinking Security in LLM Code Generation through Real-World Risk Scenarios." In *Proceedings of the 23rd Conference on Detection of Intrusions and Malware & Vulnerability Assessment (DIMVA)*, 2026.

- **Feng Wei**, Hongda Li, Ziming Zhao, and Hongxin Hu. "xNIDS: Explaining Deep Learning-based Network Intrusion Detection Systems for Active Intrusion Responses." In *Proceedings of the 32nd USENIX Security Symposium (USENIX Security)*, Anaheim, CA, USA, August 2023.
- **Feng Wei***, Hongda Li*, and Hongxin Hu. "Enabling Dynamic Network Access Control with Anomaly-based IDS and SDN." In *Proceedings of ACM International Workshop on Security in Software Defined Networks and Network Function Virtualization (SDN-NFV Security 2019)*, Richardson, TX, USA, March 2019. (*co-first author)

Invited Talks

- "xNIDS: Explaining Learning-based Network Intrusion Detection Systems for Active Intrusion Responses." VMware Research Talk, June 2021.
- "Interpreting Learning-based Network Intrusion Detection System for Active Intrusion Response." Great Lakes Security Day, November 2021.

Selected Posters

- "Explaining Learning-based Network Intrusion Detection Systems for Active Intrusion Responses." NSF/VMware SDI-CSCS Final Annual PI Meeting, 2021.
- "Dynamic Defense with Explainable Network Intrusion Detection Systems." NSF/VMware SDI-CSCS Annual PI Meeting, 2020.
- "Enabling Dynamic Network Access Control with Anomaly-based NIDS and SDN." NSF/VMware SDI-CSCS Annual PI Meeting, 2019.

Awards and Competitions

- Amazon Research Award (ARA AI for Information Security), $100,000, 2022
- USENIX Security Student Grant, 2023
- DEFCON AutoDriving Capture the Flag (CTF) Competition, 5th place (2021) and 13th place (2022)
- DJI RoboMasters Robotics Competition, Championship, 2015
- Freescale (NXP) Cup Smart Car Competition, Top 3 / 2,000 teams, 2013
- Mathematical Contest in Modeling (MCM), Meritorious Winner, 2013
- Contemporary Undergraduate Mathematical Contest in Modeling (CUMCM), 1st Prize, 2011

Professional Service

Artifact Evaluation Committee

- USENIX Security Symposium, 2022–2024
- ACM Conference on Computer and Communications Security (CCS), 2023–2024
- Annual Computer Security Applications Conference (ACSAC), 2021–2023
- ACM Conference on Data and Application Security and Privacy (CODASPY) Poster Program, 2020, 2022

Conference Reviewer

- ACM CCS, NDSS, ICML, NeurIPS, AAAI, The Web Conference (WWW), AsiaCCS, ACSAC, CODASPY

Journal Reviewer

- IEEE Transactions on Information Forensics and Security
- IEEE Transactions on Dependable and Secure Computing
- IEEE Transactions on Cloud Computing
- IEEE/ACM Transactions on Networking
- Information Systems Frontiers
- Computers & Security

Teaching Experience

- Guest Lecture: CSE 565 Computer Security, University at Buffalo, Spring 2024
- Guest Lecture: CSE 702 Machine Learning and Cybersecurity, University at Buffalo, Spring 2023

- Teaching Assistant: CPSC 8430 Deep Learning, Clemson University, Spring 2020

Technical Skills

Programming: Python, C/C++, JavaScript, LaTeX

Frameworks & Tools: TensorFlow, Keras, PyTorch, Pandas, Git, Vim, VSCode, Wireshark, Tshark

Last Updated: May 18, 2026