Jihye Choi

jihyec@nvidia.com Personal webpage Google Scholar

RESEARCH INTERESTS

Safety and security of AI agents, and agents applied to security problems:

- **Safety and security of agents**: making agents themselves safe and secure, at the model level (post-training, alignment) and at the system level (guarantees, policy constraints, verification, sandboxing).
- **Agents for security**: making agents better at security tasks by recursively self-improving the model or the agent harness, in AI security (e.g., defenses against prompt injection) and cybersecurity (e.g., discovering, verifying, and patching software vulnerabilities).

EMPLOYMENT

NVIDIA Research *May 2026 – present*
Research Scientist, Security and Privacy Research
Manager: Edward Suh

EDUCATION

University of Wisconsin–Madison *2026*
Ph.D. in Computer Sciences
Advisor: Prof. Somesh Jha

Carnegie Mellon University *2018*
M.S. in Electrical and Computer Engineering

Yonsei University *2016*
B.S., School of Integrated Technology
Early graduation with Highest Honors

EXPERIENCE

Google Cloud AI Research *May – Dec. 2025*
Research Intern with Dr. Jinsung Yoon

University of Wisconsin–Madison *2020 – 2026*
Research Assistant with Prof. Somesh Jha

Visa Research, Identity and AI Security *Summer 2022, Summer 2023*
PhD Research Intern with Dr. Yizhen Wang, Dr. Ke Wang, Dr. Rahul Rachuri

CyLab, Carnegie Mellon University *May 2017 – Mar. 2019*
Research Assistant with Prof. Lujo Bauer

T-Brain, SK Telecom *Apr. 2018 – Sep. 2018*
Research Intern with Dr. Kanghoon Lee

Vision & Learning Lab, Yonsei University *Mar. 2015 – Aug. 2016*
Research Assistant with Prof. Hyunjung Shim

PUBLICATIONS

Peer-reviewed Conference Papers

NeurIPS'26 **oral** Mihai Christodorescu, Earlence Fernandes, Ashish Hooda, Somesh Jha, Johann Rehberger, Kamalika Chaudhuri, Xiaohan Fu, Khawaja Shams, Guy Amir, **Jihye Choi**, Sarthak Choudhary, Nils Palumbo, Andrey Labunets, Nishit V. Pandya, "Agent Security is a Systems Problem," *Advances in Neural Information Processing Systems (NeurIPS)*, Dec., 2026.

COLM'26 **Jihye Choi**, Jinsung Yoon, Long T. Le, Somesh Jha, Tomas Pfister, "PolicyBank: Evolving Policy Understanding for LLM Agents," *Conference on Language Modeling (COLM)*, Oct., 2026.

ICLR'26 **Jihye Choi**, Jinsung Yoon, Jiefeng Chen, Somesh Jha, Tomas Pfister, "ATLAS: Constraints-Aware Multi-Agent Collaboration for Real-World Travel Planning," *International Conference on Learning Representations (ICLR)*, Apr., 2026.

ICLR'25 **Jihye Choi**, Jayaram Raghuram, Yixuan Li, Somesh Jha, "CONDA: Adaptive Concept Bottleneck for Foundation Models Under Distribution Shifts," *International Conference on Learning Representations (ICLR)*, Apr., 2025.

CHI'25 Christine P. Lee*, **Jihye Choi***, Bilge Mutlu, "MAP: Multi-user Personalization with Collaborative LLM-powered Agents," *ACM CHI Conference on Human Factors in Computing Systems (Late-Breaking Work)*, Apr., 2025.

MLHC'24 **Jihye Choi***, Nils Palumbo*, Prasad Chalasani, Matthew M. Engelhard, Somesh Jha, Anivarya Kumar, David Page, "MALADE: Orchestration of LLM-powered Agents with Retrieval Augmented Generation for Pharmacovigilance," *Machine Learning for Healthcare Conference (MLHC)*, Aug., 2024.

ACL'24 Neal Mangaokar*, Ashish Hooda*, **Jihye Choi**, Shreyas Chandrashekaran, Kassem Fawaz, Somesh Jha, Atul Prakash, "PRP: Propagating Universal Perturbations to Attack Large Language Model Guard-Rails," *Annual Meeting of the Association for Computational Linguistics (ACL), Long Paper*, Bangkok, Thailand, Aug., 2024.

ICML'23 **Jihye Choi**, Jayaram Raghuram, Ryan Feng, Jiefeng Chen, Somesh Jha, Atul Prakash, "Concept-based Explanations for Out-of-Distribution Detectors," *International Conference on Machine Learning (ICML)*, Jul., 2023.

ICML'23 Jiefeng Chen*, Jayaram Raghuram*, **Jihye Choi**, Xi Wu, Yingyu Liang, Somesh Jha, "Stratified Adversarial Robustness with Rejection," *International Conference on Machine Learning (ICML)*, Jul., 2023.

EI'16 **Jihye Choi**, Sungjoon Koh, Jongwoo Kwack, Yonghun Kwon, Hyunjung Shim, "Data-driven Approach to Aesthetic Enhancement," *SPIE/IS&T Electronic Imaging*, Feb., 2016.

Peer-Reviewed Journal Papers

· Clark Barrett, Brad Boyd, Elie Bursztein, Nicholas Carlini, Brad Chen, **Jihye Choi**, ..., Dawn Song, Ankur Taly, Diyi Yang, "Identifying and Mitigating the Security Risks of Generative AI," *Foundations and Trends in Privacy and Security, Vol. 6: No. 1, pp 1-52*

Peer-reviewed Workshop Papers

ICLR'26 **Jihye Choi**, Jinsung Yoon, Long T. Le, Somesh Jha, Tomas Pfister, "PolicyBank: Evolving Policy Understanding for LLM Agents," *ICLR Workshop on Lifelong Agents: Learning, Aligning, Evolving*, 2026.

ACCS'26 **Jihye Choi**, Rahul Rachuri, Ke Wang, Somesh Jha, Yizhen Wang, "SLVR: Securely Leveraging Client Validation for Robust Federated Learning," *AsiaCCS Workshop on Secure and Efficient Federated Learning*, 2026.

ICML'24 **Jihye Choi**, Jayaram Raghuram, Yixuan Li, Suman Banerjee, Somesh Jha, "Adaptive Concept Bottleneck for Foundation Models," *ICML Workshop on Foundation Models in the Wild*, 2024.

AAAI'22 **oral** Jiefeng Chen*, Jayaram Raghuram*, **Jihye Choi**, Xi Wu, Yingyu Liang, Somesh Jha, "Revisiting Adversarial Robustness of Classifiers With a Reject Option," *AAAI Workshop on Adversarial Machine Learning and Beyond*, 2022. ▷ **Best Paper Award**

Preprints

arXiv Nils Palumbo, Sarthak Choudhary, **Jihye Choi**, Guy Amir, Prasad Chalasani, Somesh Jha, "Formal Policy Enforcement for Real-World Agentic Systems," 2026.

arXiv **Jihye Choi**, Shruti Tople, Varun Chandrasekaran, Somesh Jha, "Why Train More? Effective and Efficient Membership Inference via Memorization," 2023.

arXiv Zi Wang, **Jihye Choi**, Ke Wang, Somesh Jha, "Rethinking Diversity in Deep Neural Network Testing," 2023.

arXiv Ryan Feng, Neal Mangaokar, **Jihye Choi**, Somesh Jha, Atul Prakash, "Defending Object Detectors against Patch Attacks with Out-of-Distribution Smoothing," 2022.

arXiv Kanghoon Lee*, **Jihye Choi***, Moonsu Cha, Jung-Kwon Lee, Tae Yoon Kim, "Stochastic Doubly Robust Gradient," 2018.

TEACHING EXPERIENCE

University of Wisconsin–Madison

· CS 435 Introduction to Cryptography (Fall 2021, Fall 2020, Spring 2020)
– Teaching Assistant with Prof. Somesh Jha

· CS 412 Numerical Analysis (Fall 2019)
– Teaching Assistant with Prof. Amos Ron

Carnegie Mellon University

· 18-290 Signals and Systems (Spring 2017)
– Head Teaching Assistant with Prof. Aswin C. Sankaranarayanan and Prof. Richard Stern

PROFESSIONAL ACTIVITIES

Program Committee

· AAAI, 2027

· IJCAI, 2027

· Deep Learning Security and Privacy Workshop, co-located with IEEE S&P, 2025

Workshop Organization

· Student Organizer, GenAI Risk Workshop, 2023

Conference Reviewer

· NeurIPS (2022–2024, 2026); ICLR (2026, 2027); COLM (2025, 2026); ICML (2024, 2025); CHI (2026); MLHC (2025); AISTATS (2025); CVPR (2025); ICCV (2025); SaTML (2024), external reviewer

Journal Reviewer

· IEEE TPAMI (2025)

Mentoring

· Graduate Mentor, Women in Science and Engineering (WISE), University of Wisconsin–Madison, 2025–2026

· Graduate Mentor, Korean-American Scientists and Engineers Association, Wisconsin Chapter, 2021–2022