Hyejun Jeong

Amherst, MA +1 (413) 824-1648 hjeong@umass.edu hyejunjeong.github.io linkedin.com/in/june-jeong

Research Interests

I study **security, privacy, and behavioral properties of AI agent systems**, where LLMs interact with tools and environments over multiple steps. My work analyzes **agent behavior and execution traces** to uncover **network-level information leakage**, **persuasion-induced behavioral drift**, and **persistent bias across LLM families**.

Research Experience

Research Assistant, UMass Amherst 2023–Present

- Analyzed **behavioral security risks in LLM-based AI agents**, focusing on multi-step web research and coding execution.
- Designed attacks inferring **user prompts and persona traits from domain-level browsing metadata** generated by agents.
- Studied **task-irrelevant persuasion effects on agent behavior**, showing belief-induced drift during downstream task execution.
- Developed large-scale pipelines evaluating **bias and fairness in LLMs** across 30+ models and 1M+ prompts; introduced **Bias Similarity Measurement (BSM)**.

Research Assistant, SKKU 2021–2023

- Researched defenses against poisoning and backdoor attacks in federated learning.
- Studied privacy-preserving federated learning in healthcare; co-authored multiple peer-reviewed publications.

Selected Publications

- **H. Jeong**, M. Teymoorianfard, A. Kumar, A. Houmansadr, E. Bagdasarian. "Network-Level Prompt and Trait Leakage in Local Research Agents." *USENIX Security 2026*. [Paper] [Code]
- **H. Jeong**, S. Ma, A. Houmansadr. "Bias Similarity Measurement: A Black-Box Audit of Fairness Across LLMs." *ICLR 2026*. [Paper] [Code]
- **H. Jeong**, S. Ma, A. Houmansadr. "SoK: Challenges and Opportunities in Federated Unlearning." Preprint under review. [Paper]
- **H. Jeong**, H. Son, S. Lee, J. Hyun, T.-M. Chung. "FedCC: Robust Federated Learning Against Model Poisoning Attacks." *SecureComm*, 2025. [Paper] [Code] [Slides]
- **H. Jeong**, J. An, J. Jeong. "Are You a Good Client? Client Classification in Federated Learning." *ICTC*, 2020. [Paper] [Code]
- J.H. Yoo, H.M. Son, **H. Jeong**, et al. "Personalized Federated Learning with Clustering: Non-IID HRV Data." *ICTC*, 2020. [Paper]

Additional publications in federated learning and healthcare: FDSE'22, IJWIS'22, FDSE'21, ICTC'20.

Selected Projects

- **Model Inversion on Unlearned Samples** (2024): Explored reconstructing removed samples via differences in penultimate-layer representations between original and unlearned models.
- **Federated Unlearning as Backdoor Mitigation** (2023): Investigated unlearning as a defense against backdoor attacks in FL; led literature review, experiments, and manuscript preparation. [Code]
- **Malicious Client Detection in Federated Learning** (2022): Proposed client classification method using model weight heatmaps to detect backdoors/data poisoning. Sole author of design, implementation, and write-up. [Code]

Service, Teaching & Honors

Mentorship & Service

- PhD Mentor (2023-2025) – guided undergraduates on an 11-week AI-agent security project ([Poster] presented at URV Showcase)
- Member – AISEC Lab (2025–), SPIN Group (2023–) | Reviewer – *IEEE TIFS* (2024–)

Teaching

- TA – CS690F Trustworthy AI (Fall 2025): designed and graded assignments about various AI security topics, mentored project teams
- TA – CS360 Computer & Network Security (Spring 2025): designed security-themed assignments (e.g., SHA-256 cracking, DNS spoofing), advised semester projects
- Tutor – KT Corp. AIVLE School (Feb–May 2022): coached AI-model interpretation and CS fundamentals; supported projects in ML/DL, NLP, and web-app development with Django

Education

UMass Amherst – Ph.D. in Computer Science (Exp. 2027) Advisors: A. Houmansadr, E. Bagdasaryan
SKKU, South Korea – M.S. in Computer Science (2023) Advisor: T.-M. Chung, GPA 4.5/4.5
Stony Brook Univ. – B.S. in Computer Science (2020) Specialization: Security & Privacy; Dean's List (5×)

Technical Skills

ML Frameworks: PyTorch, HuggingFace, TensorFlow, Docker, Git
Security & Privacy: AI Web Agent Attack, Persuasion, Backdoor, Federated (Un)Learning, Differential Privacy, Model Inversion
LLM Evaluation: Bias/Fairness Metrics, CKA, Cosine Similarity, SBERT, OBELS
Languages: Python, Java, C, LaTeX, JavaScript, SQL, R