Saranya Vijayakumar

saranyav@andrew.cmu.edu • Pittsburgh PA

Education

Carnegie Mellon University — Pittsburgh

Ph.D., Computer Science — *Expected graduation 2026*

M.S., Computer Science Research — *2025 via Ph.D program*

Advisors: Professors Christos Faloutsos & Matt Fredrikson. Focus: AI Safety, Alignment Security, Adversarial Robustness. Selected Coursework: Deep Learning; Grand Strategy (2023)

Harvard University — Cambridge

A.B., Joint Concentration in Computer Science & Government — *2018*

Thesis: Improving fairness and reducing harm in high-stakes algorithmic decisions
Bachelor's Advisors: Professors Cynthia Dwork & Jim Waldo.

Conference Publications

[1] **Evaluating LLM-Supported Malware Evasion: A Red Team Benchmark for Code Obfuscation and Antivirus Bypass**
Saranya Vijayakumar, Christos Faloutsos, Matt Fredrikson
Under Review — *2025*

[2] **Leveraging Large Language Models for Enhanced Membership Inference and Reidentification in Topics API Analyses**
Saranya Vijayakumar, Norman Sadeh
Under Review — *2025*

[3] **Prototype-Integrated Representation Learning for Novelty Detection**
Saranya Vijayakumar, Christos Faloutsos, Matt Fredrikson
IEEE TrustCom — *Guiyang 2025*

[4] **AICodeDetect: A Pipeline for Systematic Detection and Analysis of AI-Generated Code**
Saranya Vijayakumar, Philip Negrin, Christos Faloutsos
IEEE Mathematics and Computers in Sciences and Industry (MCSI) — *Rhodes 2025*

[5] **Mechanistically Interpreting a Transformer-based 2-SAT Solver: An Axiomatic Approach**
Nils Palumbo, Ravi Mangal, Zifan Wang, Saranya Vijayakumar, Corina Pasareanau, Somesh Jha
International Conference on Machine Learning (ICML) — *Vancouver 2025*

[6] **Aligned LLMs Are Not Aligned Browser Agents**
Priyanshu Kumar, Saranya Vijayakumar, Elaine Lau, Tu Trinh, Zifan Wang, Matt Fredrikson
The International Conference on Learning Representations (ICLR)) (Paper) — *Singapore 2025*

[7] **Grounding Neural Inference with Satisfiability Modulo Theories**
Saranya Vijayakumar, Zifan Wang, Kaiji Lu, Vijay Ganesh, Somesh Jha, Matt Fredrikson
NeurIPS (Spotlight) (Talk) — *Vancouver 2023*

[8] **CallMine: Fraud Detection and Visualization of Million-Scale Call Graphs**
Mirela Cazzolato, Saranya Vijayakumar, Meng-Chieh Lee, Namyong Park, Catalina Vajiac, Christos Faloutsos
The Conference on Information and Knowledge Management (CIKM) — *Birmingham 2023*

Workshop Publications and Conference Contributions

[1] **Through the Lens of LLMs: Unveiling Differential Privacy Challenges**
USENIX Conference on Privacy Engineering Practice and Respect (PEPR) — *Santa Clara 2024*

[2] **Anomaly Detection and Visualization of Large-Scale Call Graphs**
AAAI-23 Demonstrations Program — *Washington DC 2023*

[3] **TgraphSpot: Fast and Effective Anomaly Detection for Time-Evolving Graphs**
2022 IEEE International Conference on Big Data Industry and Government Program — *Osaka 2022*

[4] **Interpretability Through Interrogation: Fairness in the Context of Criminal Sentencing** — *2018*

[5] **Algorithmic Decision-Making** — *Harvard Political Review, 2017*

[6] **A Worldwide Survey of Encryption Products**
Bruce Schneier, Kathleen Seidel, and Saranya Vijayakumar. — *SSRN, 2015*

AI Security & Governance Speaking

[1] 17-416/17-716, AI Governance (Masters/PhD level) — Carnegie Mellon University
Guest Lecture on LLM Security and Alignment — *Spring 2025*

[2] 17-331/17-631, Information Security, Privacy, Public Policy — Carnegie Mellon University
Guest Lecture on Vulnerabilities of ML — *Fall 2024*

[3] 17-416/17-716, AI Governance (Masters/PhD level) — Carnegie Mellon University
Guest Lecture on ML Security and Privacy — *Spring 2024*

[4] 17-331/17-631, Information Security, Privacy, Public Policy — Carnegie Mellon University
Guest Lecture on ML Security and Adversarial Robustness — *Fall 2023*

[5] Dagstuhl Seminar: Machine Learning and Logical Reasoning: The New Frontier — Germany 2022

[6] CRA-WP Grad Cohort 2022 — New Orleans 2022

[7] Cylab Partners Conference — Pittsburgh 2022

[8] Alumni Committee for Harvard Women in Computer Science — 2022

Selected Fellowships and Awards

[1] Best Poster Award — New Orleans 2024

[2] GFSD Program — NSA

[3] National Defense Science & Engineering Graduate Fellowship Program — Army Research Office, 2022 – 2025

[4] Tech in the World Fellow, Partners in Health — Lima 2018

[5] The Ernst Kitzinger Prize, Lowell House — Harvard University, 2018

[6] Microsoft Scholarship, Grace Hopper Celebration of Women in Computing — Orlando 2017

Industry Experience & Security Collaborations

PNC/Fraud Detection — Pittsburgh

Collaborator — *Upcoming*

Collaboration with PNC to study financial fraud detection

IBM/Trustworthy AI — Yorktown Heights

AI Security Research Intern — *May – August 2025*

Game theoretic extension of prover-verifier games for AI system legibility using multiple specialized agentic verifiers. Developed security frameworks for AI alignment verification focusing on scalable AI safety verification. Studied under Erik Miehling and Karthikeyan Ramamurthy.

Inria/Proof techniques for security protocols (PESTO) — Nancy

Security Research Visiting Scholar — *October – November 2024*

Formal verification project: Applied formal methods to verify security properties of critical government communications infrastructure. Applied formal methods to verify cryptographic protocol implementations and identify potential vulnerabilities. Studied under Charlie Jacomme and Steve Kremer.

Mobileum/Adaptive, Intelligent and Distributed Assurance Platform (AIDA) — Braga

Security & Fraud Detection Researcher — *2021 – 2026*

Collaborated with Mobileum, a global provider of telecom analytics solutions, on industry-scale fraud detection and security research. Developed fraud detection systems protecting millions of users across 900+ operators globally. Created threat detection algorithms and security monitoring frameworks for real-time fraud prevention.

Goldman Sachs/Algorithmic Trading (GSET) — New York

Data Scientist, Electronic Trading — *2018 – 2021*

Covered quantitative hedge funds and asset managers in a client-facing data science role focusing on secure algorithmic trading systems. Performed trade cost analyses and security assessments using Python, proprietary languages, SQL, and KDB Q. Designed and implemented security-aware trading algorithms and risk assessment frameworks. Collaborated with security teams on threat modeling for high-frequency trading systems. Published research on market microstructure security and electronic trading vulnerability assessment.

Booz Allen Hamilton — Virginia Square, Herndon & Boston

Cybersecurity Research Intern — *Summer 2017*

Project: Cybersecurity for Autonomous Robotic Swarms. Created functionality for semi-autonomous navigation of ground robots in ROS using Python and C++. Conducted security research demonstrating GPS spoofing attacks against military-grade GPS-enabled robots. Implemented PCA-based anomaly detection for GPS security monitoring. Developed threat models and countermeasures for autonomous system vulnerabilities.

Digital Solutions & Policy Analysis Intern — *Winter 2017*

Performed security and privacy impact analysis on public transit systems using MBTA data. Evaluated pricing strategies and identified potential security vulnerabilities in fare collection systems. Analyzed privacy implications of surge pricing and demographic profiling in public transportation.

Beto O'Rourke for U.S. Senate — Austin

Data Scientist, Distributed Organizing — *Summer 2018*

Collaborated with data team and campaign director to create secure Python models for voter analysis while ensuring privacy protection. Developed threat assessment models for campaign security and data protection. Presented findings on digital security best practices and data privacy protocols to senior campaign leadership.

Security Research & Policy Experience

Radcliffe Institute for Advanced Study — Harvard University

Research Partner - Elite Network Analysis & Transparency — *2018*

Collaborated with investigative journalists to develop methodologies for mapping global elite networks using publicly available data. Focused on transparency in political, business, and social interrelationships with implications for security and governance.

Berkman Klein Center for Internet and Society — Harvard University

Encryption Policy Researcher — *2015*

Published comprehensive survey of worldwide encryption products under Bruce Schneier. Research influenced national cybersecurity policy and democratic governance frameworks. Research paper remained in Top 10% of SSRN downloads and was featured on Last Week Tonight with John Oliver. Analysis was used in Senate hearings on encryption regulation and influenced national cybersecurity policy discussions.

Harvard Institute of Politics — Harvard University

National Security Policy Program Director — *2017 – 2018*

Led national security policy programming and discussions. Organized speaker series on cybersecurity threats, AI governance, and national security implications of emerging technologies. Coordinated policy research initiatives on algorithmic accountability and security governance frameworks.

Security & AI Governance Teaching

17-331/631, Information, Security, Privacy & Policy (Masters level) — Carnegie Mellon University

Teaching Assistant — *Fall 2023*

Created homework assignments on ML security and privacy, developed threat modeling exercises, and led sessions on secure coding practices. Course covered applied cryptography, authentication protocols, web security, network attacks, and ML security frameworks. Designed practical exercises on vulnerability assessment and security policy development.

Future Faculty Program — Carnegie Mellon University

Participant — *2021 – 2023*

Eberly Center for Teaching Excellence & Educational Innovation. Participated in seminars aimed at helping graduate students develop and document their teaching skills in preparation for a faculty career. Completed a lesson plan review and teaching observation with Eberly experts; redesigned Rapid Prototyping syllabus; completed a teaching philosophy project

National Security & Policy Leadership

Harvard Club of New York — New York

National Security Special Interest Group, Active Member — *2018 – Present*

Regular participant in national security policy discussions and strategic analysis sessions. Contribute expertise on AI security, cybersecurity governance, and technology policy implications for national security.

Cyber Defense Club, Chair — Harvard University

Finance & Communications Chair — *2017 – 2018*

Led cybersecurity competition team that qualified for New England regional finals of the National Collegiate Cyber Defense Competition. Organized weekly security training sessions and developed incident response procedures. Managed club operations and strategic communications.

Professional Service

Program Committee, Foundations of Agentic Systems Theory	NeurIPS 2025
NeurIPS, Reviewer	2023, 2024, 2025
ICML, Reviewer	ICML 2025
ICLR, Reviewer	2024, 2025
KDD, Reviewer	KDD 2025
Peer Reviewer	Georgetown Center for Security and Emerging Technology (CSET), 2024

Selected Leadership & Service

Women in CSD, Founder — Carnegie Mellon University, 2022 – Present

Organizer of weekly programming for over 90 women and non-binary members of Computer Science Department, including security-focused career development and mentorship.

Harvard University Alumni Service — Boston & New York, 2018 - Present

Schools & Scholarships Committee: Interview Harvard College applicants annually. Participation chair for class of 2018 fifth year reunion (2023). Focus on identifying candidates with strong security and policy interests.

Technical Skills

Security Technologies: Penetration testing frameworks, vulnerability assessment tools, formal verification systems, cryptographic implementations, threat modeling, security protocol analysis, anomaly detection systems, fraud prevention algorithms.

Programming & Development: Java, Python, C, C++, R, SQL, Tensorflow, Sklearn, ROS, Git, Secure coding practices, Linux administration (bash, Apache, MySQL), KDB Q, LaTeX.

AI Safety & Alignment: Adversarial machine learning, AI alignment verification, robustness testing, long-term AI safety research, scalable oversight methodologies, LLM safety assessment, differential privacy, membership inference attacks, robustness testing.

Natural languages: English, Tamil, Spanish (*working proficiency*), Japanese (*limited working proficiency*).