AI Institute for Agent-based Cyber Threat Intelligence and Operation

https://action.ucsb.edu

# ACTION Year 1 Annual Report: Summary

## Reporting period: June 2023 to May 2024 (Published November 2026)

The NSF AI Institute for Agent-based Cyber Threat Intelligence and Operation (ACTION) develops AI-enabled security agents that work with human defenders across the full cybersecurity life cycle: assessing vulnerabilities, detecting attacks, attributing them, and responding and recovering. ACTION is led by UC Santa Barbara with 10 partner universities and is supported by the U.S. National Science Foundation, with additional support from the Department of Homeland Security and IBM.

**Research.** In its first year, ACTION launched 25 research activities: three research threads in each of four foundational AI thrusts and four cybersecurity thrusts, plus the Institute's shared testbed, GATE. ACTION researchers published more than 35 papers, with about as many more forthcoming, including four at CVPR 2024, six at ICLR 2024, and 16 at ICML 2024. Highlights include:

- SAIN, developed at Purdue and published at USENIX Security 2024, which detected a simulated cyberattack on a power plant in GATE's fictional city of New Esperanza, alerted human analysts, and kept the plant in a safe state.
- Research on jailbreak attacks against large language models, which is informing a shared knowledge base for ACTION's intelligent agents.
- DeepCapa, which identifies the malicious capabilities of malware from a snapshot of its memory and explains its decisions by pointing to the responsible code.
- A transfer learning approach for malware detection with limited labeled data, developed by Purdue researchers with IBM and explored with Cisco for commercial use.

**Shared testbed.** ACTION built the first version of GATE, a cloud-based testbed modeled on the fictional city of New Esperanza. Researchers demonstrated a network traversal attack designed at Georgia Tech and simulated cyber-physical systems developed at Purdue.

**Competitions and community.** Dr. Radha Poovendran and Dr. Bo Li organized the Trojan Detection Challenge (LLM Edition) at NeurIPS 2023, and a UCSB team led by Dr. Giovanni Vigna and Dr. Christopher Kruegel entered DARPA's AI Cyber Challenge. ACTION launched a seminar series with distinguished lectures and student talks, and sponsored UCSB's 2023 Responsible Machine Learning Summit.

**Education and workforce.** ACTION built programs from high school through graduate school:

- The 20th International Capture the Flag (iCTF), redesigned around AI topics, drew 38 undergraduate and 10 high school teams. Six ACTION institutions built 26 new challenges.
- Saturday workshops for high school students through UCSB's School for Scientific Thought led to outreach at Santa Barbara High School, supported by a 2024 UCSB Chancellor's Faculty Outreach Grant.
- ACTION's first summer research experience received 113 applications for 12 places, and selected eight undergraduate and four graduate students to begin in June 2024.

**Impact and engagement.** Co-PI Dr. Christopher Kruegel was named a 2023 ACM Fellow. Dr. Stuart Russell published on the regulation of advanced AI agents in *Science* and gave written testimony to the U.S. Senate AI Insight Forum. Dr. Xiaoming Huo and Dr. Bin Yu organized a session on Statistical Methods in Cybersecurity for the 2024 Joint Statistical Meetings, and ACTION researchers presented their work to a visiting team of DARPA program specialists.

**Looking ahead.** In May 2024, ACTION revised its Strategic and Implementation Plan to detail how its research thrusts integrate and how progress will be measured. Year 2 priorities include deploying intelligent agents in GATE, launching the postdoctoral program, developing new AI and security courses, and establishing a student-led cybersecurity club at Santa Barbara High School.

*The full report is available at https://action.ucsb.edu/knowledge-transfer. ACTION's current Strategic & Implementation Plans are available by request.*
*Please direct any questions to action@cs.ucsb.edu.*

This material is based upon work supported by the U. S. National Science Foundation under grant no. 2229876 and is supported in part by funds provided by the U. S. National Science Foundation, by the Department of Homeland Security, and by IBM.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U. S. National Science Foundation or its federal agency and industry partners.