AI Institute for Agent-based Cyber Threat Intelligence and Operation,

a National AI Research Institute

University of California, Santa Barbara
Purdue University
Georgia Institute of Technology
University of California, Berkeley

Giovanni Vigna, Director
Tim Robinson, Managing Director

Year 3 Annual Report
Reporting period: June 2025 to May 2026

For public release.
This report describes activities through May 31, 2026.

Cooperative agreement: 2229876

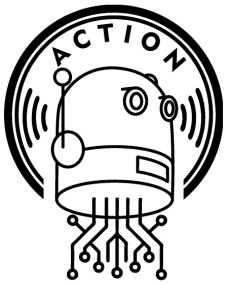


Table of Contents

Project Summary **4**

TABLE 1 – PRINCIPAL ORGANIZATIONS OF THE INSTITUTE 5

TABLE 2 - LIST OF THE LEADERSHIP TEAM 5

TABLE 3 - LIST OF MAJOR INSTITUTE ACTIVITY PARTICIPANTS 7

TABLE 4 - LIST OF OTHER COLLABORATORS AND PARTNERS 11

TABLE 5 - ADVISORY BOARDS 11

Executive Summary **13**

Narrative **18**

1.0 Vision 18

2.0 Strategic Plan 19

3.0 Essential Functions of the Institute 20

3.1 Research 20

3.1.1 Foundational AI research 21

Thrust AI-1: Learning and Reasoning with Domain Knowledge 21

Thrust AI-2: Human-Agent Interaction 23

Thrust AI-3: Multi-Agent Collaboration 24

Thrust AI-4: Strategic Gaming and Tactical Planning. 26

Globally Accessible Testing Environment (GATE) 28

Thrust SEC-1: Intelligent Agents for Risk/Vulnerability Assessment 30

Thrust SEC-2: Intelligent Agents for Detection. 32

Thrust SEC-3: AI-driven Cyber Threat Intelligence 33

Thrust SEC-4: Intelligent Agents for Response and Recovery. 35

3.1.2 Other Cross-Institute research 37

3.1.3 Benefits of use-inspired research 41

3.2 Nurture and grow the next generation of talent 42

3.2.1 University and college education 42

3.2.2 Pre-college education 48

3.2.3 Workforce development 48

3.3 Broadening participation and diversity 48

3.4 Multidisciplinary integration and community building 51

3.5 Multi-organizational synergies/achievements 52

3.6 Knowledge transfer 53

TABLE 6 INTELLECTUAL PROPERTY 57

TABLE 7 - TECHNOLOGY TRANSFER 58

TABLE 8 - START-UP FIRMS 59

3.7 Impact of the Institute as a nexus point for collaborative efforts 60

4.0 Evaluation and Assessment 63

4.1 Evaluation Report | Executive Summary 63

4.2 ACTION Response to Evaluation Report 67
5.0 Infrastructure and Management 72
5.1. Configuration and Leadership 72
5.2. Organizational Chart 73
Individual Roles and Responsibilities 73
5.3. Management Systems 75
5.4. Ethics 76
References 77
Appendix A: Glossary and Acronyms 83
Appendix B: Highlights of Significant Achievement and Impact 88
Highlight 1: Lifelong Detection and Learning of New Threats 88
Highlight 2: Adaptive Design of AI Assistance to Enhance Human Decision Making 89
Highlight 3: Learning with Regret Minimization for Cybersecurity Applications 90
Highlight 4: Leto: Reinforcement Learning to Optimize Intrusion Response Against… 91
Highlight 5: Shield Agent: Unified and Knowledge-Enabled Guardrail Agents to Protect… 92
Highlight 6: Klleem: Multi-Level, Targeted Vulnerability Analysis in Large Networks 93
Highlight 7: iCTF 2026 "Battle of the Agents" and the Agentish Platform 95
Highlight 8: CVE-Genie: Automated End-to-End Vulnerability Reproduction 96
Highlight 9: ARTIPHISHELL: Autonomous Cyber Reasoning at Machine Speed 97
Highlight 10: PurpCorn-PLAN: Winning the Amazon Nova AI Challenge 99
Highlight 11: TOUCAN: Supercharging Tool-Calling AI Agents 100
Appendix C: List of Institute Projects 102
Appendix D: Project Updates (June 2025-May 2026) 113

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.

Project Summary

Overview:
The sheer size, complexity, and rapidly changing nature of modern computer systems, as well as the constantly evolving tactics, techniques, and procedures (TTPs) of adversarial attacks, make it very challenging to properly assess these systems' security posture and respond to and recover from attacks. Although many existing cyber-defense solutions incorporate artificial intelligence (AI) and machine learning (ML) for defense, they tend to automate specific manual analysis tasks and typically focus only on parts of a system or individual steps in the security life cycle. The AI Institute for Agent-based Cyber Threat Intelligence and Operation (ACTION) initiates a new approach to cybersecurity, in which AI-enabled intelligent security agents cooperate with humans across the cyber-defense life-cycle tasks to jointly improve the security posture of complex computer systems over time. New foundational AI research will develop an AI stack that provides novel fundamental primitives for lifelong, continual reasoning and learning with domain knowledge, human-agent interaction that supports effective teaming, multi-agent collaboration that is adversary-aware, and strategic gaming that provides provable guarantees about the effectiveness of a course of action. Intelligent agents will leverage the capabilities provided by this new AI stack to support novel threat and vulnerability assessment approaches that take into account the context and dependencies of a system, the detection of evasive adversaries, the attribution of complex attacks, and the planning of effective response and recovery activities. The ACTION Institute will also innovate education and workforce development, and create opportunities for industry collaboration while functioning as a nexus for the AI and cybersecurity communities.

Intellectual Merit:
By using novel knowledge-modeling and logic reasoning techniques to support the collaboration of intelligent security agents, ACTION will provide the breakthroughs necessary for an agent-driven autonomous process that continuously improves the security of computer systems.

Broader Impacts:
The ACTION Institute's work will increase the resilience of critical infrastructure against attacks carried out by sophisticated actors, possibly backed by nation-states. Sensitive data and critical services will be protected, saving billions of dollars and, in some cases, human lives. The Institute will leverage its infrastructure and network of industry and community partners to create educational opportunities in AI and cybersecurity with the explicit goal of engaging diverse groups of students and professionals. In particular, its large-scale and challenge-based learning activities will ensure reach and impact well beyond the core partner institutions.

Participant Tables

TABLE 1 – PRINCIPAL ORGANIZATIONS OF THE INSTITUTE

Name of Institution	Role	City	State/Country
University of California, Santa Barbara	Lead Institution	Santa Barbara	CA, USA
Purdue University	Domestic Partner Institution (Core)	West Lafayette	IN, USA
Georgia Institute of Technology	Domestic Partner Institution (Core)	Atlanta	GA, USA
University of California, Berkeley	Domestic Partner Institution (Core)	Berkeley	CA, USA
University of Washington	Domestic Partner Institution	Seattle	WA, USA
University of Illinois at Urbana-Champaign	Domestic Partner Institution	Champaign	IL, USA
University of Illinois at Chicago	Domestic Partner Institution	Chicago	IL, USA
University of Chicago	Domestic Partner Institution	Chicago	IL, USA
University of Virginia	Domestic Partner Institution	Charlottesville	VA, USA
Rutgers University	Domestic Partner Institution	New Brunswick	NJ, USA
Norfolk State University	Domestic Partner Institution	Norfolk	VA, USA
Western Washington University	Domestic Partner Institution	Bellingham	WA, USA

Table 1 updates: (Western Washington University officially joined ACTION this year as a subawardee).

TABLE 2 - LIST OF THE LEADERSHIP TEAM

Position Title	Name	Department	Organization
Director	Giovanni Vigna	Computer Science	University of California, Santa Barbara
Co-PI	Dongyan Xu	Computer Science	Purdue University
Co-PI	Wenke Lee	Computer Science	Georgia Institute of

			Technology
Co-PI	Dawn Song	Electrical Engineering and Computer Science	University of California, Berkeley
Senior Personnel	Chris Kruegel	Computer Science	University of California, Santa Barbara
Senior Personnel	João Hespanha	Electrical and Computer Engineering	University of California, Santa Barbara
Senior Personnel	Ambuj Singh	Computer Science	University of California, Santa Barbara
Senior Personnel	Xiaoming Huo	Data Science and Statistics	Georgia Institute of Technology
Senior Personnel	David Evans	Computer Science	University of Virginia
Senior Personnel	Ninghui Li	Computer Science	Purdue University
Senior Personnel	Elisa Bertino	Computer Science	Purdue University
Senior Personnel	Berkay Celik	Computer Science	Purdue University
Senior Personnel	Ming Yin	Computer Science	Purdue University
Senior Personnel	Yung-Hsiang Lu	Electrical and Computer Engineering	Purdue University
Senior Personnel	Jie Gao	Computer Science	Rutgers University
Senior Personnel	Gang Wang	Computer Science	University of Illinois at Urbana-Champaign
Senior Personnel	Bo Li	Computer Science	University of Illinois at Urbana-Champaign
Senior Personnel	Nick Feamster	Computer Science	University of Chicago
Senior Personnel	Bing Liu	Computer Science	University of Illinois at Chicago
Senior Personnel	Radha Poovendran	Electrical and Computer Engineering	University of Washington
Senior Personnel	Sewoong Oh	Computer Science and Engineering	University of Washington
Senior Personnel	David Wagner	Electrical Engineering and Computer Science	University of California, Berkeley
Senior Personnel	Pieter Abbeel	Electrical Engineering and Computer Science	University of California, Berkeley
Senior Personnel	Bin Yu	Statistics	University of California, Berkeley
Senior Personnel	Stuart Russell	Electrical Engineering and Computer Science	University of California, Berkeley
Senior Personnel	Tonya Fields	Computer Science	Norfolk State University

Senior Personnel	Mary Ann Hoppa	Computer Science	Norfolk State University
Senior Personnel	Jules Chenou	Computer Science	Norfolk State University
Senior Personnel	Bhaskar Ramasubramanian	Computer Science	Western Washington University
Managing Director	Tim Robinson	Institute Staff	University of California, Santa Barbara
Software Engineer	Kyle Quinlan	Institute Staff	University of California, Santa Barbara

Table 2 updates:

- Bhaskar Ramasubramanian joined ACTION as the lead faculty from Western Washington University. Dr. Ramasubramanian will contribute both to educational objectives and to research.
- Kyle Quinlan joined ACTION as a software engineer in February 2026. Onboarding concluded by March.

TABLE 3 - LIST OF MAJOR INSTITUTE ACTIVITY PARTICIPANTS

Thrust AI-1: Learning and Reasoning with Domain Knowledge			
Position Title	Name	Department	Institution
Activity Leader	Bo Li	Computer Science	University of Illinois at Urbana-Champaign
Faculty Researcher	Bin Yu	Statistics	University of California, Berkeley
Faculty Researcher	Dawn Song	Electrical Engineering and Computer Science	University of California, Berkeley
Faculty Researcher	Stuart Russell	Electrical Engineering and Computer Science	University of California, Berkeley
Faculty Researcher	Bing Liu	Computer Science	University of Illinois at Chicago
Faculty Researcher	Ambuj Singh	Computer Science	University of California, Santa Barbara
Faculty Researcher	Sewoong Oh	Computer Science and Engineering	University of Washington
Thrust AI-2: Human-Agent Interaction			
Position Title	Name	Department	Institution
Activity Leader	Ming Yin	Computer Science	Purdue University

Faculty Researcher	Gang Wang	Computer Science	University of Illinois at Urbana-Champaign
Faculty Researcher	David Wagner	Electrical Engineering and Computer Science	University of California, Berkeley
Faculty Researcher	Radha Poovendran	Electrical and Computer Engineering	University of Washington
Faculty Researcher	Bing Liu	Computer Science	University of Illinois at Chicago
Faculty Researcher	Ambuj Singh	Computer Science	University of California, Santa Barbara
Thrust AI-3: Multi-Agent Collaboration			
Position Title	Name	Department	Institution
Activity Leader	Jie Gao	Computer Science	Rutgers University
Faculty Researcher	Xiaoming Huo	Data Science and Statistics	Georgia Institute of Technology
Faculty Researcher	Bin Yu	Statistics	University of California, Berkeley
Faculty Researcher	João Hespanha	Electrical and Computer Engineering	University of California, Santa Barbara
Faculty Researcher	Ninghui Li	Computer Science	Purdue University
Faculty Researcher	Jules Chenou	Computer Science	Norfolk State University
Faculty Researcher	Tonya Fields	Computer Science	Norfolk State University
Faculty Researcher	Ambuj Singh	Computer Science	University of California, Santa Barbara
Faculty Researcher	Pieter Abbeel	Electrical Engineering and Computer Science	University of California, Berkeley
Faculty Researcher	Wenke Lee	Computer Science	Georgia Institute of Technology
Thrust AI-4: Strategic Gaming and Tactical Planning			
Position Title	Name	Department	Institution
Activity Leader	Radha Poovendran	Electrical and Computer Engineering	University of Washington
Faculty Researcher	João Hespanha	Electrical and Computer Engineering	University of California, Santa Barbara

Faculty Researcher	Wenke Lee	Computer Science	Georgia Institute of Technology
Thrust SEC-1: Intelligent Agents for Threat/Vulnerability Assessment			
Position Title	Name	Department	Institution
Activity Leader	Gang Wang	Computer Science	University of Illinois at Urbana-Champaign
Faculty Researcher	David Wagner	Electrical Engineering and Computer Science	University of California, Berkeley
Faculty Researcher	Dawn Song	Electrical Engineering and Computer Science	University of California, Berkeley
Faculty Researcher	Giovanni Vigna	Computer Science	University of California, Santa Barbara
Faculty Researcher	Chris Kruegel	Computer Science	University of California, Santa Barbara
Faculty Researcher	Bo Li	Computer Science	University of Chicago
Faculty Researcher	Berkay Celik	Computer Science	Purdue University
Faculty Researcher	Radha Poovendran	Electrical and Computer Engineering	University of Washington
Faculty Researcher	Dongyan Xu	Computer Science	Purdue University
Faculty Researcher	Elisa Bertino	Computer Science	Purdue University
Thrust SEC-2: Intelligent Agents for Detection			
Position Title	Name	Department	Institution
Activity Leader	David Evans	Computer Science	University of Virginia
Faculty Researcher	Dongyan Xu	Computer Science	Purdue University
Faculty Researcher	Sewoong Oh	Computer Science and Engineering	University of Washington
Faculty Researcher	Elisa Bertino	Computer Science	Purdue University
Faculty Researcher	Berkay Celik	Computer Science	Purdue University
Faculty Researcher	Giovanni Vigna	Computer Science	University of California, Santa Barbara

Faculty Researcher	Chris Kruegel	Computer Science	University of California, Santa Barbara
Thrust SEC-3: Intelligent Agents for Attribution			
Position Title	Name	Department	Institution
Activity Leader	Nick Feamster	Computer Science	University of Chicago
Faculty Researcher	Dongyan Xu	Computer Science	Purdue University
Faculty Researcher	David Evans	Computer Science	University of Virginia
Faculty Researcher	Giovanni Vigna	Computer Science	University of California, Santa Barbara
Faculty Researcher	Yung-Hsiang Lu	Electrical and Computer Engineering	Purdue University
Thrust SEC-4: Intelligent Agents for Response and Recovery			
Position Title	Name	Department	Institution
Activity Leader	Elisa Bertino	Computer Science	Purdue University
Faculty Researcher	Berkay Celik	Computer Science	Purdue University
Faculty Researcher	Dongyan Xu	Computer Science	Purdue University
Faculty Researcher	Wenke Lee	Computer Science	Georgia Institute of Technology
Faculty Researcher	Giovanni Vigna	Computer Science	University of California, Santa Barbara
Globally Accessible Testing Environment			
Position Title	Name	Department	Institution
Activity Leader	Berkay Celik	Computer Science	Purdue University
Software Engineer	Kyle Quinlan	Institute Staff	University of California, Santa Barbara
Faculty Researcher	Dongyan Xu	Computer Science	Purdue University
Faculty Researcher	Wenke Lee	Computer Science	Georgia Institute of Technology
Faculty Researcher	Giovanni Vigna	Computer Science	University of California, Santa Barbara

Table 3 updates: None

TABLE 4 - LIST OF OTHER COLLABORATORS AND PARTNERS

Name of Institution / organization / partner	City	State / Country	Role
University of California, Center for Evaluation and Assessment, Tarek Azzam	Santa Barbara	CA, USA	Evaluator
National Institute of Information and Communications Technology (NICT) , Akira Tanaka	Koganei	Tokyo, Japan	Research Collaborator

Table 4 updates: A researcher from NICT has joined ACTION for a short research exchange. He is funded entirely by his home institution.

TABLE 5 - ADVISORY BOARDS

External Advisory Board			
Name	Title	Organization (Department or Division)	Institution or Firm
Snehal Antani	CEO		Horizon3.ai
Igor Tsyganskiy	Chief Strategy Officer	Security	Microsoft
Wen Masters	VP	Cyber Technologies	MITRE
Tom Gillis	SVP/GM	Security	Cisco
Latanya Sweeney	Professor of the Practice of Government and Technology		Harvard University
Mary Ann Blair	Chief Information Security Officer		Carnegie Mellon University
TBN			
TBN			
Student Advisory Council			

Name	Title	Organization (Department or Division)	Institution or Firm
Chase (Xi) Jiang	PhD Student	Computer Science	University of Chicago
Fengqing Jiang	PhD Student	Electrical and Computer Engineering	University of Washington
Zhun Wang	PhD Student	Electrical Engineering and Computer Science	University of California, Berkeley
Saastha Vasan (Co-chair)	PhD Student	Computer Science	University of California, Santa Barbara
Leyan Pan	PhD Student	Computer Science	Georgia Institute of Technology
Doguhan Yeke	PhD Student	Computer Science	Purdue University
Daniel Baumgartner	PhD Student	Mathematics	Rutgers University
Hadjer Benkraouda (Co-chair)	PhD Student	Computer Science	University of Illinois at Urbana-Champaign
Minjun (Elena) Long	PhD Student	Computer Science	University of Virginia
Saleh Momeni	PhD Student	Computer Science	University of Illinois, Chicago
Ralph Watson-Quartey	Post grad	Computer Science	Norfolk State University

Table 5 updates:

- The SAC roster has changed due to the graduating class of 2025. We anticipate another change of membership in the next reporting period.
- This year, UIC and Norfolk both maintained representation on the SAC.
- Western Washington University does not have student representation at this time. This may change as WWU becomes more involved in ACTION. Their funding, however, has been limited to Summer Support for Dr. Ramasubramanian, for his contributions to ACTION's educational objectives.
- The SAC will consider changes to their membership in Year 4.

Executive Summary

Narrative:

The AI Institute for Agent-based Cyber Threat Intelligence and Operation (ACTION) completed its third year of operation with significant progress across all eight research thrusts, three education initiatives, and a growing portfolio of knowledge transfer activities. Research highlights include provably optimal game-theoretic algorithms for cyber-physical system defense, an award-recognized program in adaptive human-AI decision-making, and lifelong learning agents that autonomously incorporate new threats without human retraining. Beyond the eight research thrusts, ACTION released a number of benchmarks and tools, including TOUCAN and CVE-GENIE, that have attracted immediate interest from IBM, Cisco, Intel, NIST, and others. ACTION faculty also advanced GATE, the institute's globally accessible cyber-physical testbed, with new embodied AI attack and defense agents and a least-privilege framework that blocked 98.8% of attacks while maintaining full agent utility. ACTION's Google Scholar profile now reflects over 300 manuscripts, 8,000+ citations, and an h-index of 44 [1], and institute researchers demonstrated real-world impact through their participation in two high-profile competition achievements: a first-place finish and $250,000 prize at the Amazon Nova AI Challenge and a top-five finish at the DARPA AI Cyber Challenge.

On the education and knowledge transfer front, ACTION created courses for agentic programming, received a $2M NSF education supplement connecting four universities in a joint AI+Cybersecurity curriculum effort, and hosted CTF training camps for high school students, undergraduates, and U.S. veterans. The institute deepened its government engagement through monthly meetings with DHS and CISA, a briefing to the NITRD Interagency Working Group, and an interview with GAO science writers for a forthcoming Congressional brief on AI policy. Internationally, ACTION contributed to the Center for Digital Trust summit at EPFL and joined a global working group drafting policy recommendations on AI adoption and control. ACTION believes these activities position us as an emerging nexus connecting government agencies, industry partners, international research institutions, and the next generation of the cybersecurity workforce.

This summary highlights top achievements across the eleven categories defined by NSF's Essential Functions of the Institute framework.

Highlight Achievements in the Reporting Period (June 2025 - May 2026):

i. Foundational AI research (see Section 3.1.1)
   a. Developed region-aware regret minimization algorithms with provable optimality guarantees that adapt across the full spectrum of adversarial variability, extended to defend cyber-physical systems against stealthy sensor and actuator attacks (Thrust AI-4).
   b. Produced an adaptive AI ensemble with a "Rational Routing" mechanism that dynamically switches between trust-building and performance-boosting models to improve human-AI decision-making; published at AAAI 2026 and recognized with a best paper honorable mention at CHI 2026 (Thrust AI-2) [2, 3].
   c. Advanced lifelong learning algorithms that autonomously detect and incorporate new threats without human-labeled data, transforming static post-deployment models into continuously improving systems (Thrust AI-1).

ii. Other Cross-Institute research (see Section 3.1.1 and 3.1.2)
   a. GATE advanced on two fronts: a least-privilege framework for LLM agents blocked 98.8% of attacks while preserving full task utility, and a new multi-agent framework for embodied robotic AI automatically identified adversarial failure scenarios and demonstrated that state-of-the-art safety defenses in robotics can be bypassed.
   b. Released TOUCAN, the largest publicly available dataset of its kind, synthesizing 1.5 million data points from nearly 500 real-world Model Context Protocols to accelerate development of tool-using AI agents; developed in collaboration with IBM.
   c. Introduced CVE-GENIE, an automated multi-agent system that reproduces verifiable exploits for approximately 51% of real-world vulnerabilities at an average cost of $2.77 per CVE, generating significant interest from CISCO, Intel, NIST, Red Hat, and others

iii. Benefits of use-inspired research (see Section 3.1.3)
   a. ACTION researchers produced six benchmarks spanning the institute's research thrusts, from LLM safety and graph reasoning to embodied AI security and web agent evaluation, providing the community with rigorous, operationally grounded tools for procurement, red-teaming, and system evaluation.

b. Team PurpCorn-PLAN (UIUC), including ACTION researcher Hadjer Benkraouda and Professor Gang Wang, won first place and a $250,000 prize at the Amazon Nova AI Challenge, a head-to-head tournament testing the safety of AI coding models under adversarial conditions.
c. Team Shellphish (UCSB), founded by ACTION Director Giovanni Vigna, finished 5th at the DARPA AI Cyber Challenge final at DEF CON 33, with the highest patching success rate (>95%) among finalists and the best telemetry of the seven competing teams.

iv. University and college education (see Section 3.2.1)
a. Hosted summer research experiences at UCSB and Rutgers, engaging seven undergraduate and graduate students from the California State University system and other institutions; one resulting project ("COVERTLENS") was accepted and presented at a NeurIPS 2025 workshop.
b. Launched zero2hero, a CTF-based educational platform developed jointly by UCSB and Norfolk State University, introducing 40 new CTF challenges with AI-powered tutors and walkthroughs for students with no prior cybersecurity knowledge.
c. Received a $2M NSF education supplement to develop a composable, modular curriculum at the intersection of AI and cybersecurity, spanning six proposed courses targeting undergraduate, graduate, and professional audiences; led primarily by the University of Washington and Georgia Tech, broadening ACTION's educational reach well beyond UCSB.

v. Pre-college education (see Section 3.2.2)
a. Norfolk State University hosted its first ACTION-sponsored CTF training camps for 25 high school students and seven teachers from school systems across the Hampton Roads region of Virginia.
b. Professor Kruegel hosted a high school student from San Jose, CA for a summer research experience at UCSB, contributing to an active ACTION research project called REDACTBENCH, alongside undergraduate and postdoctoral colleagues.
c. Launched the iCTF 2026 "Battle of the Agents" competition with agentic challenges (Agentish) for the first time, attracting high school teams from across the U.S.

vi. Workforce development (see Section 3.2.3)
a. NSU's ACTION-sponsored CTF night camp was attended by 22 students, including five U.S. veterans, providing hands-on cybersecurity training through CTF exercises; 2025 summer intern Anoushka Sawant was awarded a fellowship with the NSF-funded National Community College

Cybersecurity (NCyTE) Center to train the next generation of cybersecurity educators.

b. ACTION faculty began adapting existing courses at Georgia Tech, UW, and WWU to align with the proposed composable AI+Cybersecurity curriculum, including Introduction to Information Security, ML4Cyber, and TinyML.
c. Developed an AI-powered teacher/tutor chatbot (Poovendran) and a novel graduate course (AVATAR: Autonomous Vulnerability Analysis, Triaging, And Repair) as early deliverables of the NSF education supplement.

vii. Broadening participation and diversity (see Section 3.3)
   a. Summer research programs at UCSB and Rutgers, drawing many students from the California State University system, and summer CTF camps at Norfolk State University, a Historically Black University, engaged students from Hispanic-Serving Institutions and Historically Black Colleges and Universities.
   b. The iCTF 2026 high school division attracted teams from diverse school districts, with prize support enabling participation from underrepresented schools.
   c. Norfolk State University's CTF training camps specifically targeted students and veterans from underserved communities in the Hampton Roads region, including participants from Chesapeake, Portsmouth, and Suffolk.

viii. Multidisciplinary integration and community building (see Section 3.4)
   a. ACTION's Student Advisory Council played an active role in vetting and selecting summer research participants, strengthening cross-campus student connections and leadership development.
   b. Managing Director Robinson participated in the 3rd Annual Center for Digital Trust (C4DT) Summit at EPFL, Switzerland, contributing to a global working group on AI policy and education and co-authoring a peer-reviewed study on project management practices across 25 NSF National AI Research Institutes.
   c. Increased cross-campus graduate student interaction, driven by new cohort transitions, produced measurable gains in community cohesion and collaborative research momentum across ACTION campuses.

ix. Multi-organizational synergies/achievements (see Section 3.5)
   a. Received a $2M NSF education supplement, connecting UCSB, UW, Georgia Tech, and Western Washington University in a joint curriculum development effort targeting the AI+Cybersecurity education gap.

b. The supplement opened direct lines of communication with NSF's Faculty Development Summer Institute, NSA's Centers for Academic Excellence in Cybersecurity (497 campuses), and CISA's CYBER.ORG platform.
c. ACTION's education supplement catalyzed new meetings with IBM's Talent Development team, exploring integration of ACTION educational tools with IBM's SkillsBuild programs at HBCUs and other universities.

x. Knowledge transfer (see Section 3.6)
a. Director Vigna delivered the opening keynote at ACM CCS 2025 in Taipei, a keynote at Amazon's Bot Buster's Summit, and a briefing to the NITRD Cyber Security and Information Assurance Interagency Working Group; Professor Lee delivered the U.S. keynote at the NSF/CSA Singapore Workshop on Cybersecurity Innovations; Director Vigna was interviewed by GAO science writers for a Congressional brief on AI policy.
b. CVE-GENIE and TOUCAN generated significant industry traction: CVE-GENIE drew coverage from Reversing Labs and outreach from CISCO, Intel, NIST, ChainGuard, Red Hat, and others; TOUCAN was highlighted in an IBM blog post and made publicly available on GitHub.
c. ACTION meets monthly with DHS Science and Technology and CISA scientists to share research achievements and provide early access to new manuscripts, and through IBM has been introduced to CoSAI (Coalition for Secure AI) as a new industry community for engagement.

xi. Impact of the institute as a nexus point for collaborative efforts (See Section 3.7)
a. ACTION's Google Scholar profile tracks over 300 manuscripts, 8,000+ citations, an h-index of 44, and an i10-index of 147, providing a quantifiable measure of scholarly influence; ACTION represented all NSF AI Institutes at the 2025 SCSP AI+Expo in Washington, D.C., engaging more than 15,000 government, industry, and academic leaders.
b. Participated in the C4DT Summit at EPFL, contributing to a global working group on AI policy and a collaborative manuscript, led by Warwick University and the Turing Institute, identifying a complete AI stack for national AI adoption and control policy.
c. Positioned ACTION as an emerging nexus for cybersecurity education by building relationships with DHS/CISA (new AP Cybersecurity course), NSA's CAE program (497 campuses), IBM SkillsBuild, and local CTE programs near ACTION campuses, connecting government, industry, and academia around a shared workforce development agenda.

Narrative

1.0 Vision

The Institute for Agent-based Cyber Threat Intelligence and Operation (ACTION) will initiate a revolutionary approach to cybersecurity in which AI-enabled intelligent security agents cooperate with humans across cyber-defense life-cycle tasks to jointly improve the security posture of complex computer systems over time.

We define an intelligent security agent as an entity that performs one or more cybersecurity functions, with reasoning, learning, and collaboration (human-agent and inter-agent) capabilities. Intelligent security agents follow a new paradigm of continuous, lifelong learning both in autonomy and in collaboration with human experts, supported by a shared knowledge bank and an integrated AI stack that provides novel fundamental primitives for reasoning and learning with domain knowledge, human-agent interaction, multi-agent collaboration, and strategic gaming and tactical planning.

Over time, these intelligent security agents will improve their domain knowledge and their ability to perform security tasks such that they will be increasingly robust and effective in the face of changes in the adversaries' modes of operation, capable of composing defense strategies and tactical plans in the presence of uncertainty, collaborative with each other and with humans for mutually complementary teaming, and adaptive to unfamiliar and novel attacks.

The goal of the ACTION Institute is to realize the above vision, via integrated research that synergizes novel foundational AI concepts and use-inspired approaches to cybersecurity. Research results will expand the foundations of AI and enable high automation, scalability, and cost-effectiveness in both proactive and reactive cyber-defense. Throughout the life of the Institute, the use-inspired research will highlight new challenges in knowledge representation and learning, agent/agent and agent/human interaction, and game theory and planning, which, in turn, will inspire new AI foundational research, realizing a virtuous circle between theoretical AI and its applications.

The ACTION Institute's research efforts will be complemented by innovation in education (K-12, undergraduate, graduate, and post-graduate). On the smaller scale, we

will immerse select students in targeted research exchanges at partner institutions. More broadly, we will host security exercises, such as Capture the Flag competitions, that have an underlying educational impact and will generate classroom tools to improve security training at the high school and undergraduate level. Finally, we will create opportunities for online learning and improved workforce development tools, which will be made available at no cost to the community.

ACTION aims to work with the U.S. National Science Foundation, with IBM Corporation, and with the U.S. Department of Homeland Security Science and Technology Division to broaden these agencies' reach and potential impact. Through monthly meetings with DHS, where we provide briefings on new research, we also engage with the Cybersecurity and Infrastructure Security Agency (CISA). As an Institute, we have been introduced to these stakeholders' research efforts, as well as their own plans and visions for education and workforce development.

In this regard, ACTION's vision is to foster collaboration between its member campuses, strategic partners, and other organizations existing both within and beyond the United States. We have ambitious plans to deliver research results able to integrate with, and improve upon, the state of the art for cyber defense. Through the development, implementation, and utilization of our testbed infrastructure and our AI Stack, ACTION will become a global nexus between the AI and cybersecurity communities by 2028.

2.0 Strategic Plan

With only a few exceptions, we followed our Year 3 Strategic Implementation Plan (S&IPv3), guiding activities from June 2025 - May 2026. Exceptions to the plan were

- The addition of educational activities as a result of supplemental funding from NSF
- The design of a novel platform that launched the first ever "agentic" Capture the Flag postponed the annual iCTF from December 2025 to March 2026
- The frequency of meetings with stakeholders (DHS and IBM) increased, while communication with our External Advisory Board decreased
- The submission of this annual report occurred later than planned

These disruptions were minor considering the increased collaboration and activity inspired by the changes written into the Year 3 strategic plan. ACTION continues to

maintain a structure supported by four AI-focused research thrusts and four thrusts that correspond to the cybersecurity lifecycle. The eight thrusts are connected by various distinct threads that make up the outline of a novel AI stack. Our managerial structure did not change over the course of Year 3, but was instead strengthened by the collaborative requirements of the various mini-projects undertaken by the faculty researchers and their students. Appendix C lists all Year 3 research projects, and Appendix D summarizes progress in each research thrust.

ACTION's S&IPv4, which will direct our activities from June 2026 - May 2027, is published as a separate document. Outlined in this strategic plan are 18 mini-projects that demonstrate tangible collaborations between faculty from separate campuses and promise deliverables that advance both foundational AI and use-inspired goals. We drafted this version of the plan in May 2026 with a narrowed overarching scenario meant to inspire a greater sense of focus and urgency among our researchers.

Beyond the narrowed focus for Year 4, ACTION's S&IPv4 suggests two additional changes to our overall strategic plan. The first is an increased output of our own educational materials along with increased efforts to connect the educational missions of our various stakeholders. An education supplement received in Summer 2025 opened a direct line of communication with the organizers of NSF's Faculty Development Summer Institute and their program on AI in Cybersecurity Education and those associated with the Centers for Academic Excellence in Cybersecurity, a community with funding from NSA. These connections, coupled with IBM's SkillsBuild program [4] and CISA's CYBER.ORG [5], have brought to light ideas about how ACTION can strengthen its position as a nexus for Cybersecurity in Education over the next 12-18 months.

3.0 Essential Functions of the Institute

3.1 Research

During our third program year, ACTION progressed research along all eight research thrusts and GATE through 20 distinct research/technology projects. Because our strategic plan prescribes a 12-month base period of performance for each of these investigations, we often refer to them as "mini-projects."

We introduced the Year 3 projects in our S&IPv3 and provided abstracts for each of them throughout the previous two quarterly reports. Appendix C provides a full list of Year 3 ACTION research projects and associated personnel. Appendix D summarizes the goals, research threads, and achievements of each research thrust, GATE, and our education and outreach programs for the reporting year. While we identify faculty researchers as the project leads, we would also like to acknowledge the implied participation of the many graduate students advised by the co-investigators.
One area requiring candid acknowledgment is the AI stack. We completed significant stack-relevant work across all eight thrusts, from lifelong learning, adversary-aware multi-agent collaboration, game-theoretic planning, and secure agent design. However, as of the end of Year 3, ACTION has not yet produced a formal definition or architecture document for the stack. Defining the stack in concrete terms is a primary deliverable for Year 4, and the S&IPv4 mini-projects have been selected with this integration goal in mind. We expect to have a substantive update at the Year 4 site visit.

In the next section, we describe ACTION's successful Year 3 projects, sorted by their primary research thrusts. To demonstrate the interconnectedness of our research, we take care to identify where the mini-projects impact or advance the efforts of other research thrust.

ACTION researchers also undertook a number of use-inspired research projects beyond the mini-projects set in the S&IPv3. These efforts reflect the breadth and impact of ACTION and are highlighted in Section 3.1.2 Other Cross-Institute Research.

3.1.1 Foundational AI research

This section discusses ACTION research. For each of the eight research thrusts and GATE, we highlight successes and challenges of our 20 mini-projects. We also use these projects to connect thrusts to one another. Appendix D summarizes progress in each thrust. We include discussions on our education and outreach, on broadening participation and knowledge sharing in later sections of the report.

Thrust AI-1: Learning and Reasoning with Domain Knowledge

Which projects have been successful?

Thrust AI-1 focuses on equipping intelligent agents with the ability to learn and reason continuously, at scale, and in dynamic environments. We saw three successful

mini-projects under this thrust not only advance this work, but also complement one another.

Ambuj Singh and Jie Gao's project develops time-dependent representations that allow agents to detect and predict evolving cyber threats in real time.

David Evans and Sewoong Oh provide the methods needed to encode and interpret complex temporal data streams, giving agents a richer understanding of how threats unfold over time.

Bing Liu, Xiaoming Huo and Wenke Lee ensure that agents can autonomously incorporate new knowledge and shed erroneous knowledge without relying on human engineers to retrain them from scratch.

Together, these projects build toward a unified vision of agents that not only understand the current threat landscape but continuously refine that understanding as the world changes around them.

How do project successes enable other projects to move forward?

The contributions of these three projects enable aspects of ACTION's other seven research thrusts. Dynamic graph modeling directly strengthens multi-agent collaboration (AI-3) by enabling agents to share and act on evolving threat intelligence, and supports threat detection (SEC-2) by providing richer models of how attacks propagate through interconnected systems. LLM-based time series reasoning similarly feeds into SEC-2 as the project moves toward incorporating real cybersecurity data. Lifelong learning is perhaps the most broadly impactful: its ability to autonomously detect and adapt to new threats informs human-agent interaction (AI-2) by reducing the burden on human experts, supports response and recovery (SEC-4) through unlearning mechanisms that correct erroneous knowledge post-deployment, and lays the foundational reasoning infrastructure that strategic gaming (AI-4), vulnerability assessment (SEC-1), attribution (SEC-3), and intrusion detection (SEC-2) all depend upon.

Delays, setbacks, and plans to address these in the coming period(s)?

There were no delays or setbacks with these three projects. Pending the acceptance of recently submitted manuscripts [6-12], Singh and Gao's Dynamic Graphs project can be

considered complete. Liu and Lee's Detection project can also be considered complete with the submission of three papers [13-15].

As Evans and Oh's Time Series Reasoning project [16] moves into its next phase, it will advance the efforts of Thrust SEC-2: Threat Detection. The goal is to incorporate cybersecurity data generated by other ACTION projects in our experiments.

Major new directions for the proposed year(s).

We have no major direction changes for Thrust AI-1 planned for the upcoming year. The thrust will continue to support AI capabilities that will ultimately allow agents to continuously learn from their environment and reason about necessary actions to take.

Other considerations as appropriate.

N/A

Thrust AI-2: Human-Agent Interaction

Which projects have been successful?

Thrust AI-2 focuses on enabling meaningful, effective interaction between humans and intelligent security agents, and ensuring that AI assistance adapts to individual needs, builds appropriate trust, and supports sound collective decision-making. In support of these goals, **Ming Yin and Ambuj Singh** address the individual level, developing frameworks that personalize how AI recommendations are presented and challenged, balancing trust-building with performance optimization through mechanisms like adaptive model ensembles and counterfactual critical thinking. **Yin and Singh** then scale these ideas to the team level, exploring how LLM-powered agents can help groups maintain shared knowledge, aggregate distributed information, and delegate tasks effectively. Together, the two projects span much of the spectrum of human-agent interaction, from the individual analyst deciding whether to trust an alert, to the cross-organizational team coordinating a response to a widespread attack.

How do project successes enable other projects to move forward?

The insights generated by these two projects extend naturally across other ACTION research thrusts. Most directly, both projects inform threat detection (SEC-2) by making

human analysts less susceptible to over-reliance, alert fatigue, and adversarial manipulation, all of which are exploitable situations. The adaptive ensemble and counterfactual reasoning frameworks also strengthen threat assessment (SEC-1) by helping analysts reason more carefully about complex, ambiguous threat signals. The collective decision-making work ties into multi-agent collaboration (AI-3) by establishing principles for how human and AI agents can coordinate under uncertainty, and into strategic gaming (AI-4) by informing how teams can reach consensus and respond decisively under adversarial pressure. The lifelong learning infrastructure of AI-1 underpins both projects, as agents must continuously update their models of individual and group behavior, while the response and recovery thrust (SEC-4) benefits directly from the team coordination mechanisms developed here.

Delays, setbacks, and plans to address these in the coming period(s)?

There were no delays or setbacks with AI-2 research. Pending publication of accomplished work [17-20], we consider these projects to be complete.

Major new directions for the proposed year(s).

We have no major direction changes for Thrust AI-2 planned for the upcoming year. In the overarching scenario presented in S&IPv4, the defense agents will be required to interact with operators of the plants' infrastructure to capture priorities and directives. These capabilities are aligned with the initial goals set forth by the thrust.

Other considerations as appropriate.

N/A

Thrust AI-3: Multi-Agent Collaboration

Which projects have been successful?

Thrust AI-3 focuses on enabling intelligent agents to collaborate effectively in environments where some agents may be malicious, information may be corrupted, and trust cannot be assumed. ACTION faculty successfully approached this thrust from two complementary angles. **Ninghui Li and Jie Gao** tackled the core technical problem directly, developing methods that enforce security policies on what LLM-based agents disclose and enabling agents to reason soundly even when the information they receive

may have been compromised. **Bo Li and Bin Yu** approached the problem using the backdrop of healthcare, where the consequences of unreliable agent behavior can be severe. Through their Green Shielding framework, **Li and Yu** reveal how even routine, non-adversarial variations in user inputs can systematically shift LLM behavior in clinically meaningful ways, providing a user-centric foundation for safer agent deployment.

Together, the four faculty advanced a shared vision of multi-agent systems that are not only capable but trustworthy, secure, and robust to both adversarial manipulation and benign variation.

How do project successes enable other projects to move forward?

The contributions of the AI-3 mini-projects connect deeply to other ACTION thrusts. The adversary-aware reasoning techniques developed in AI-3 feed directly into the foundational learning and reasoning infrastructure of AI-1, and the security policy enforcement mechanisms are essential inputs to strategic gaming and threat modeling in AI-4. On the security side, the ability to reason under corrupted information directly strengthens threat assessment (SEC-1) and intrusion detection (SEC-2), where agents must make sound decisions despite incomplete or manipulated data.

The Green Shielding framework's insights into how input variation affects agent behavior inform both vulnerability assessment (SEC-1) and CTI attribution (SEC-3), where prompt-level factors could be exploited by adversaries. The MedAgent architecture, orchestrating specialized agents with structured knowledge access and safe interaction, also serves as a model for the kind of human-agent teaming envisioned in AI-2, and its integration with GATE ties directly into the response and recovery infrastructure of SEC-4.

Delays, setbacks, and plans to address these in the coming period(s)?

There were no significant setbacks or delays to these projects. Pending publication of submitted manuscripts [6, 21-23], we consider them complete.

Major new directions for the proposed year(s).

Thrust AI-3 will continue to support agents' abilities to collaboratively support the security process in novel, unexpected ways. While this capability does not represent a

new direction for the research, it does imply a reliance on SEC thrusts, specifically SEC-3, which is working to allow agents to share threat intelligence.

Other considerations as appropriate.

N/A

Thrust AI-4: Strategic Gaming and Tactical Planning.

Which projects have been successful?

Thrust AI-4 focuses on developing game-theoretic approaches that model adversarial behavior, provide provable guarantees on detection effectiveness, and enable agents to plan and adapt strategically in dynamic cyber environments. AI-4 was supported by three successful projects this reporting period that form a deeply complementary research program.

João Hespanha and Radha Poovendran provide the theoretical backbone with their work on learning with regret minimization. They develop learning algorithms with provable performance guarantees that adapt across the full spectrum of adversarial variability, from stable, slowly evolving threats to rapidly shifting, fully adaptive adversaries, while also extending these ideas to the defense of cyber-physical systems.

The projects led by **Wenke Lee and Christopher Kruegel** and by **Dongyan Xu and Radha Poovendran** are especially impactful, listed both under AI-4 and SEC-2. Lee and Kruegel's multi-level vulnerability analysis project (Klleem) operationalizes strategic thinking at the code level, using LLM-guided Monte Carlo simulation to identify the highest-risk attack paths and then directing fuzzing and symbolic execution precisely where they are most needed.

Xu and Poovendran's low-and-slow APT detection-evasion project completes the picture by rigorously characterizing how sophisticated adversaries game current detection systems, revealing specific structural weaknesses that the other two projects' algorithms must be hardened against.

How do project successes enable other projects to move forward?

As noted, the three AI-4 projects advance a unified vision of AI agents that not only respond to known threats but reason strategically about how adversaries think, plan, and adapt.

The projects also radiate across the other ACTION thrusts. The regret minimization algorithms directly support response and recovery (SEC-4) by providing provably robust decision-making under adversarial uncertainty, and inform multi-agent collaboration (AI-3) by establishing how agents can coordinate effectively even when facing deceptive opponents.

The Klleem vulnerability analysis framework feeds directly into threat assessment (SEC-1) by enabling targeted, scalable discovery of exploitable weaknesses, and its LLM-guided reasoning connects to the foundational learning infrastructure of AI-1.

The APT evasion study serves as what the team calls "adversarial grounding" for the entire ACTION stack. The study's findings on how attackers mimic benign profiles directly inform the development of more resilient intrusion detection systems (SEC-2), more accurate adversary attribution (SEC-3), and more realistic training scenarios for strategic gaming across all of AI-4. Together, these projects also inform human-agent interaction (AI-2) by clarifying the kinds of adversarial blind spots that human analysts are most likely to fall prey to, and that AI assistance must be designed to counteract.

Delays, setbacks, and plans to address these in the coming period(s)?

There were no significant setbacks or delays to these projects. Pending publication of submitted manuscripts [24-27], we consider them complete.

Major new directions for the proposed year(s).

We have no major direction changes for Thrust AI-4 planned for the upcoming year. The thrust will continue game-theoretic research into how agents can devise dynamic, reasoning-based collaborative plans that reflect the current status of the available infrastructure.

Other considerations as appropriate.

N/A

Globally Accessible Testing Environment (GATE)

Which projects have been successful?

We made significant progress on two parallel fronts for GATE during Year 3. In one direction, the GATE team developed a suite of agents for evaluating security in embodied robotic AI systems. They built a multi-agent framework that automatically identifies failure scenarios for embodied robotic agents by discovering physically plausible, adversarially introduced semantic changes that alter a robot's visual-semantic reasoning and cause it to execute wrong or incomplete actions. Complementing this, they also developed an attack agent that bypasses state-of-the-art safety defenses in embodied AI. This activity demonstrates that robots can be manipulated into executing harmful actions despite Asimov-style rule enforcement.

We also produced two benchmark agents: one that evaluates existing attack and defense agents across diverse scenarios measuring both security and utility; and another that systematically evaluates a robot's ability to correctly refrain from action. We discuss these benchmarks in greater detail below, in Section 3.1.3. These contributions led to three manuscripts submitted in May 2026, targeting NDSS and NeurIPS [28, 29].

For a second direction, the GATE team began developing agentic AI components for reverse engineering (RE) tasks without human intervention.

Additionally, work on a systematic context-aware least-privilege framework for LLM agents produced one of GATE's most concrete deployable results: a framework that maps every possible agent interaction into six categories of excessive permissions and enforces strict least-privilege rules across all of them, blocking 98.8% of attacks across 13 of 15 attack categories while maintaining full agent task utility. This GATE contribution is further discussed below in Section 3.1.2.

How do project successes enable other projects to move forward?

GATE's Year 3 achievements connect directly to nearly every ACTION research thrust, functioning as the institute's primary evaluation and validation infrastructure. The multi-agent framework for embodied AI vulnerability discovery advances the goals of SEC-1 (vulnerability assessment) and SEC-2 (threat detection) by providing a principled, reproducible environment for stress-testing AI systems against adversarial manipulation. The attack and defense agents developed for GATE feed directly into the benchmark contributions of *RoboJailBench* [28] and *The Yes-Man Syndrome* [29], both of

which emerged from GATE-adjacent work and provide the field with new evaluation standards for embodied AI security. The least-privilege framework supports SEC-4 (response and recovery) by hardening the agent architectures that autonomous response systems depend upon, and its methodology informs AI-3's adversary-aware multi-agent collaboration agenda. The reverse engineering agents connect to SEC-1's binary analysis work and complement the LLM-based RE project led by **Gang Wang and Berkay Celik**.

We will have GATE serve as the deployment and validation platform for ACTION's AI stack. As ACTION's agents mature across all eight thrusts, GATE will be the environment where they will be tested, compared, and evaluated under realistic adversarial conditions.

Delays, setbacks, and plans to address these in the coming period(s)?

GATE's development suffered a significant setback this year with the departure of a leading software engineer. Replacing him required two rounds of applications and six months of recruitment, during which GATE development fell primarily to Dr. Celik and his graduate students. A new engineer, **Kyle Quinlan**, joined in February 2026, and with his onboarding, GATE development is expected to resume at full speed entering Year 4.

The delay affected the planned integration of synthetic log pipelines from the NetDiffusion and NetSSM [30] projects led by **Nick Feamster and Wenke Lee**, but not too significantly. We anticipate the project to be completed in Year 4.

Major new directions for the proposed year(s).

In Year 4, the GATE team will continue in two primary directions. The first allows ACTION to expand into physical AI and robotic system security, with a specific focus on LLM-controlled robots. We plan to deploy the attack, defense, and benchmark agents developed in Year 3 directly into GATE's infrastructure, expanding its cyber-physical evaluation capabilities and making these tools available to practitioners and educators who lack access to physical robots for security testing. This positions GATE as a shared community resource for embodied AI security research, which is an area of growing urgency as autonomous systems become more prevalent in defense, infrastructure, and industrial settings.

The second direction extends our work into reverse engineering and binary analysis, further integrating the Codex-based reverse engineering agent for solving CTF challenges and the enhanced skill-based RE tool developed within GATE that structures reverse engineering into modular skills for controlled experimentation and benchmarking. This effort focuses on integration and evaluation rather than model development, and connects GATE's infrastructure directly to ACTION's SEC-1 vulnerability assessment agenda.

Thrust SEC-1: Intelligent Agents for Risk/Vulnerability Assessment

Which projects have been successful?

The overall goal of Thrust SEC-1 is to develop knowledge-driven, context-aware approaches to vulnerability analysis and threat assessment approaches that move beyond traditional static scanning and head toward AI-powered systems that reason about risk, prioritize analysis, and understand the full scope of a system's attack surface. Our two SEC-1 projects undertaken in Year 3 advance toward this goal.

Gang Wang and Berkay Celik's binary reverse engineering project approaches threat assessment from the analyst's perspective, building an LLM-powered agent that mimics human reasoning to perform end-to-end reverse engineering. It identifies vulnerabilities, generates hypotheses, and verifies results autonomously across a growing suite of real-world challenges. Critically, it also investigates the cognitive biases that LLMs themselves are susceptible to, ensuring that the agents we build are not themselves easily fooled.

Also highlighted above in Thrust AI-4, **Wenke Lee and Christopher Kruegel's** vulnerability analysis project approaches the same problem from a system-wide perspective. Using LLM-guided Monte Carlo simulation to identify the highest-risk attack paths across a network, the system then directs deep code analysis precisely where it matters most. Together, the two projects cover complementary layers of threat assessment. One focuses on understanding individual binaries and their weaknesses. The other focuses on mapping risk across large, interconnected systems. Both leverage LLM reasoning to dramatically accelerate and deepen the analysis that human analysts alone could not perform at scale.

How do project successes enable other projects to move forward?

The SEC-1 projects also connect naturally to other ACTION research. Both projects draw heavily on the foundational learning and reasoning capabilities of AI-1, as their LLM-based agents must reason carefully about code structure, attack semantics, and vulnerability context. The multi-agent architectures underlying both systems tie directly into AI-3, with the reverse engineering project explicitly planning for a planner agent to coordinate the RE pipeline.

The Klleem framework's risk-prioritization logic reflects the game-theoretic thinking of AI-4, identifying the attack paths that a strategic adversary is most likely to exploit. On the security side, the vulnerability findings generated by both projects feed directly into intrusion detection (SEC-2) by identifying the weaknesses that detection systems must be tuned to catch, and into attribution (SEC-3) by characterizing the techniques and tools that adversaries are likely to use.

The reverse engineering agent's ability to analyze binaries autonomously also supports response and recovery (SEC-4) by accelerating post-incident forensic analysis, while both projects' outputs inform the human-agent interaction designs of AI-2 by clarifying where human analyst judgment remains essential and where it can be safely delegated to AI.

Delays, setbacks, and plans to address these in the coming period(s)?

We have no significant delays or setbacks to report. These projects will support future research in this area and may be expanded upon.

Major new directions for the proposed year(s).

As outlined in S&IPv4, Thrust SEC-1 will continue to support capabilities for defense agents that are constantly assessing the security posture of their assigned infrastructure. There are no major direction changes.

Other considerations as appropriate.

N/A

Thrust SEC-2: Intelligent Agents for Detection.

Which projects have been successful?

Thrust SEC-2 focuses on detecting events across multiple domains and abstraction levels. Specifically, we aim to develop approaches that provide robust training of detection agents, the handling of threat evolution and evasion, and novel detection approaches that incorporate uncertainty measures and multi-agent collaboration. SEC-2 is arguably ACTION's most active research thrust at this time, with the greatest number of dedicated mini projects.

Christopher Kruegel and David Evans's malware classification project tackled one of the most persistent practical problems in deployed detection systems: false positives driven by policy discrepancy. Here, **Kruegel and Evans** develop ensemble classifiers that rapidly adapt to the specific operational context of each deployment environment.

David Wagner, Dawn Song, and Bo Li's *Shield Agent* addressed a different but equally urgent vulnerability. The agentic systems that power modern security operations are themselves susceptible to prompt injection, data poisoning, and tool manipulation. *Shield Agent* builds a unified guardrail framework to protect them.

The *Leto* project, led by **Gang Wang and Christopher Kruegel**, moved beyond detection entirely, using reinforcement learning to automatically optimize how networks respond to intrusions, demonstrating in a real-world NCSA deployment that it can block hundreds of thousands of malicious connections with zero false positives.

Dongyan Xu, Radha Poovendran and Sewoong Oh led a low-and-slow APT evasion project that essentially serves as the adversarial foundation for the entire SEC-2 thrust. Here, the faculty systematically demonstrate how current provenance-based detection systems can be gamed by patient adversaries who mimic benign behavior. This finding directly informs what other ACTION projects must be hardened against.

Finally, **Sewoong Oh and Elisa Bertino** addressed the data problem underlying all detection systems, investigating how transfer learning can be made resistant to evasion, poisoning, and private data extraction attacks.

How do project successes enable other projects to move forward?

SEC-2's projects impact the entire ACTION ecosystem. The adversarial grounding provided by the APT evasion study informs strategic gaming (AI-4) by characterizing

the attack strategies that game-theoretic models must anticipate, and attribution (SEC-3) by profiling how sophisticated adversaries operate during the low-and-slow phase.

Shield Agent's privilege-separation and safety-reasoning techniques connect directly to multi-agent collaboration (AI-3) and to the secure LLM agent designs being developed in SEC-4.

Leto's RL-based response optimization ties naturally into response and recovery (SEC-4) and benefits from the game-theoretic methodology of AI-4 in handling evasive attackers. The context-adaptive malware classifier's post-deployment adaptation framework advances the lifelong learning goals of AI-1, while its SOC feedback modeling informs the human-agent interaction designs of AI-2 by clarifying how analyst judgment should be incorporated into detection workflows.

The transfer learning project's analysis of zeroth-order optimization attacks informs vulnerability assessment (SEC-1) and multi-agent security (AI-3), since understanding the capabilities of black-box attack methods is essential for hardening the agentic systems that ACTION is building across all thrusts.

Delays, setbacks, and plans to address these in the coming period(s)?

We report no setbacks or delays on the SEC-2 projects.

Major new directions for the proposed year(s).

The scenario driving our S&IPv4 requires agents to identify attacks on networks, endpoints, and services. Accordingly, Thrust SEC-2's major goals will not change.

Other considerations as appropriate.

N/A

Thrust SEC-3: AI-driven Cyber Threat Intelligence

Which projects have been successful?

Thrust SEC-3 sits at the intelligence layer of the ACTION cybersecurity stack. Its mission is to understand attacks deeply enough to attribute them to specific actors,

uncover adversarial collaboration, and anticipate future campaigns. Achieving this requires solving three interconnected problems: knowing what intelligence practitioners actually need, generating sufficient high-quality data to train attribution models, and building classifiers sophisticated enough to distinguish between malware families and characterize adversary behavior with precision.

Giovanni Vigna and David Evans CTI research-practice gap project provides the strategic foundation, systematically mapping where current academic CTI research falls short of real-world practitioner needs and rigorously evaluating where LLMs can and cannot be trusted to automate the CTI generation and sharing pipeline. Finding that LLMs struggle with false positives, policy compliance, and overestimating attacker capabilities, the project defines the problem space that other SEC-3 projects must address.

Nick Feamster and Wenke Lee's project on synthetic network attack log generation tackles the data scarcity problem that underlies all CTI research, using diffusion-based and state-space generative models to produce high-fidelity, labeled attack traces that would otherwise be impossible to obtain at scale in privacy-sensitive operational environments.

Dongyan Xu and Giovanni Vigna's malware classification project completes the picture by enriching the MABEL benchmark [31] with over 1,000 attributes per sample and building ensemble classifiers that combine multiple feature modalities. In this way, they have developed the richest possible representation of malware behavior and enabled the kind of nuanced family attribution and behavior characterization that supports meaningful adversary profiling.

Together, the three projects address the full CTI pipeline: understanding what practitioners need, generating the data to train attribution models, and building the models themselves.

How do project successes enable other projects to move forward?

The CTI research-practice gap findings directly inform human-agent interaction (AI-2) by clarifying where human analyst judgment must remain central in the CTI workflow and where AI automation is safe to deploy.

The synthetic log generation project is perhaps the most broadly enabling: high-fidelity synthetic attack data supports intrusion detection (SEC-2) by providing training data for ML-based detectors, informs vulnerability assessment (SEC-1) by simulating realistic

attack scenarios, and feeds into the GATE testbed infrastructure that underlies all of SEC-4's response and recovery experimentation.

The malware classification ensemble's multi-modal, LLM-enhanced feature representations draw directly on the foundational learning and reasoning advances of AI-1, and its dynamic model selection strategy reflects the kind of adaptive, context-aware reasoning that AI-3's multi-agent collaboration framework is designed to support.

More broadly, the adversary behavioral profiles generated by SEC-3 serve as essential inputs to the game-theoretic models of AI-4, which require accurate characterizations of attacker tactics and objectives in order to compute meaningful strategic responses, and to the response and recovery planning of SEC-4, which must anticipate attacker behavior in order to devise effective countermeasures.

Delays, setbacks, and plans to address these in the coming period(s)?

The SoK paper project suffered some setbacks, first due to reviewer comments and then due to formatting issues upon submission. We have finalized and submitted a revision. Pending acceptance, we regard this project to be complete.

With the updates to MABEL, we also regard this stage of the project to be complete.

Major new directions for the proposed year(s).

With Thrust SEC-3, agents will continue to produce, share, and consume threat intelligence about attack instances.

Other considerations as appropriate.

N/A

Thrust SEC-4: Intelligent Agents for Response and Recovery.

Which projects have been successful?

With Thrust SEC-4, we aim to empower intelligent agents to respond to and recover from cyberattacks effectively, leveraging human-agent and inter-agent collaboration to devise adaptive strategies that make systems increasingly resilient over time. With the

thrust undergoing considerable revisioning in Year 2, we had only one project that saw significant progress: **Wenke Lee and David Wagner's** project on an adaptable defense for LLM-based computer use agents. This project addressed prompt injection, which is what industry practitioners identify as the number one threat to AI agent security. The connection to response and recovery is preemptive. Rather than patching existing agent architectures, the team developed a fundamentally more secure agent design pattern, *Secure ReAct,* that preserves the flexibility of current agentic workflows while providing formal guarantees against prompt injection attacks. Parallel work has begun extending these protections to vision-language models.

How do project successes enable other projects to move forward?

Since SEC-4 is somewhat "all inclusive," requiring agents to constantly assess, detect, gather information, and respond while recovering from an incident, GATE is a natural testbed for the thrust's success. Updates to GATE provide the infrastructure that makes rigorous evaluation of response and recovery strategies possible, while simultaneously advancing the science of agent security in its own right.

Delays, setbacks, and plans to address these in the coming period(s)?

There haven't been any significant setbacks to SEC-4's specific agenda.

Major new directions for the proposed year(s).

Lee and Wagner are currently wrapping up implementation of their preliminary prototypes/baselines for *Secure ReAct*. Once each of these defenses are implemented, they will evaluate them on these datasets and consider future steps. Subsequent evaluations are not expected to impair advancement along the thrust, nor do they appear to require a significant change of plans.

Other considerations as appropriate.

N/A

3.1.2 Other Cross-Institute research

In addition to the 20 mini-projects listed in Appendix C, ACTION researchers made a number of other advancements over the past year that advance our institute's general research agenda.

As a follow-up to MAGPIE [32, 33], the fully automated method for generating LLM training data we introduced in our previous quarterly report and presented during our Year 2 site visit, we present two additional AI tools for security research: TOUCAN [34] and CVE-GENIE [35].

TOUCAN is a massive new dataset designed to supercharge the development of AI assistants, specifically LLM agents, that can use various software tools to complete complex tasks. Researchers lack high-quality, real-world data showing an agent learning to use a tool, particularly when a task requires multiple steps or switching between different tools. To bridge this gap, TOUCAN synthesized 1.5 million data points from nearly 500 authentic Model Context Protocols (MCPs), which are real-world environments used for training and testing AI. Unlike older, less diverse datasets, TOUCAN's data is realistic, complex, and includes the actual execution logs of the tools being used, making it effective at teaching AI agents how to plan, interact, and course-correct when using software.

ACTION researchers led by **Radha Poovendran** developed a pipeline to ensure the high quality and diversity of the TOUCAN dataset. First, they used multiple language models to generate a broad array of realistic tool-use questions. These initial questions were then passed through a rigorous quality filter to discard any poor examples. Next, three different "teacher" models were used to generate the correct, step-by-step solutions (called trajectories), simulating how a top-performing AI agent would interact with the tools to solve the problem. The team also incorporated methods to simulate multi-turn conversations and introduce variety into the tasks, making the dataset more challenging and reflective of real-world use. By making TOUCAN the largest publicly available dataset of its kind, the authors aim to accelerate the creation of highly capable, open-source AI agents that can seamlessly use software tools to automate tasks across a wide range of fields.

TOUCAN realizes research from Thrusts AI-1 and AI-3 and will play a role in SEC-4 in future reporting periods.

CVE-GENIE is an automated, multi-agent system designed to solve a significant challenge in cybersecurity: the manual, time-consuming process of turning basic

vulnerability descriptions (i.e. CVE entries) into verifiable, working exploits. When the ArXiv paper was mentioned on LinkedIn, CVE-GENIE received immediate attention. CVE-GENIE advances the research threads of Thrust SEC-3: Intelligent Agents for Intelligence and Attribution. The ACTION leads are **Giovanni Vigna and Christopher Kruegel**; but the work is led by a student from Boston University and also has collaborators from Arizona State University.

To summarize the problem solved by CVE-GENIE, high-quality security datasets are crucial for evaluating security tools and training AI, but security datasets are scarce because of the difficulties and cost of gathering technical details, reconstructing the exact vulnerable software environment, and writing the attack code. CVE-GENIE automates this entire pipeline, performing four key steps: i. gathering initial vulnerability data; ii. automatically rebuilding the specific vulnerable software version; iii. generating the exploit code; and then iv. independently verifying that the attack actually works. This robust, collaborative design, which the authors found was necessary after single-model attempts failed, marks a major advancement in creating reliable ground-truth data for the field.

In a comprehensive evaluation, CVE-GENIE successfully reproduced verifiable exploits for approximately 51% (428 out of 841) of real-world vulnerabilities published in 2024 and 2025, covering a broad range of programming languages and vulnerability types. Critically, the system works at an extremely low average cost of just $2.77 per vulnerability, demonstrating high efficiency and scalability. By providing a low-cost, automated method to generate reliable and reproducible security benchmarks, CVE-GENIE creates a valuable resource for the security community, enabling better evaluation of security testing tools, improving patching efforts, and allowing for standardized assessment of artificial intelligence's capabilities in the security domain.

ROXAS. Elisa Bertino's group has designed and experimentally evaluated an approach to determine whether explanations concerning anomalies in cellular network control plane support recommendations by LLM for mitigations of potential security attacks. Our work is motivated by the complexity of next-generation networks which translates to a broad attack surface, increasingly hard to monitor and protect. In the ongoing cybersecurity arms race, as attackers exploit Artificial Intelligence (AI) to design new threats, defenders must operate at the same level. While AI-based anomaly detection has shown great promise, its interpretability is often hindered by the black box nature of modern models.

Explainable AI (XAI) methods, starting from feature attribution, can address this challenge by providing insights into the model's decision-making process. Yet, it

remains unclear to what extent XAI can help analysts interpret alerts and guide mitigation actions. In this work, we have explored the use of large language models (LLMs) as an additional interpretive layer in an anomaly detection pipeline. We designed ROXAS (Reasoning Over eXplained AnomalieS), a methodology that combines anomaly detection via an XGBoost regressor trained solely on benign data, logic-based feature attribution for correct interpretation of alerts, and LLM-based guidance to move toward actionable mitigation assistance. We evaluated the LLM outputs against the expert-curated MITRE FiGHT database. The results show alignment with best practices in 5G network defenses. The results of this work are reported in a paper at IEEE International Conference on Computing, Networking and Communications [36].

ChronoMem. Bertino's group has also designed and experimentally evaluated a version control and semantic rollback approach for large language model (LLM) agent memory. Our work is motivated by the need of mitigating attacks/errors that poison the agent's memory (SEC-4). LLM agents increasingly rely on long-term memory to support multi-session interaction and personalization. However, existing agent memory systems are designed around forward-only evolution (continuously accumulating, consolidating, and overwriting knowledge) with no principled mechanism to inspect, version, or revert prior states. This makes agents brittle under corrections, concept drift, and memory corruption, particularly after the agent has already been exposed to subsequent information. To address such issues, we have designed ChronoMem, a semantic version-control layer for agentic memory, integrated into a production-ready open-source agent framework (Agent Development Kit by Google). ChronoMem commits whole-memory snapshots at each memory write, maintains structured version histories, and supports natural-language rollback requests by mapping "undo" intents to concrete historical versions via hybrid lexical–semantic retrieval, rank fusion, and reranking. We further introduce a post-exposure evaluation protocol that tests whether an agent can behave counterfactually after rollback, answering queries and summarizing history as if future updates never occurred. On long-horizon conversational benchmarks augmented with evolving memory states and rollback tasks, ChronoMem substantially improves rollback-consistent question answering and history summarization relative to prompt-only and retrieval-only baselines, while achieving strong recovery on semantic version selection. To our knowledge, ChronoMem is the first open-source system and benchmark for systematic, semantic, global memory rollback in LLM agents. A paper is being prepared for submission to a conference.

A Systematic Context-aware Least-Privilege Framework for LLM Agents. AI agents powered by large language models (LLMs) can behave unpredictably, sometimes by

accident, and sometimes because an attacker has manipulated them. Current systems try to limit what these agents are allowed to do based on the task at hand, but they often leave gaps that attackers can exploit. **Wenke Lee's** approach to this problem starts by systematically mapping out every way an AI agent can interact with a system, identifying six categories of potentially excessive permissions. We then built a framework that enforces strict "least privilege" rules across all six categories, meaning each agent only has exactly the permissions it needs for the specific step it is currently performing, nothing more.

In testing, our framework blocked every attack in one standard benchmark (AgentDojo [37]) without interfering with the agent's ability to do its job. In a broader test across 25 tasks and 15 types of attacks, including eight attacks specifically designed to evade our defenses, it blocked 98.8% of attacks while still letting the agent complete its work successfully.

Targeted Adversarial Evasion Attack against Vertical Federated Learning. Vertical Federated Learning (VFL) enables multiple parties to collaboratively train models using different features of the same samples. Most existing VFL attacks focus on the training phase (e.g., label inference, backdoor insertion, gradient manipulation), while largely neglecting inference-time evasion. Our work analyzes the robustness of VFL in object classification tasks, focusing on how a malicious party can craft adversarial perturbations to its local features to mislead global predictions toward a target class. Prior work often relies on traditional adversarial examples and overlooks feature-level interactions across participants, which introduce unique vulnerabilities in VFL. Consequently, existing perturbations fail to effectively suppress benign features contributions. Joint work by **Elisa Bertino and Sewoong Oh** explores how an adversary can override other participants by identifying and perturbing high-influence local features, exposing structural weaknesses in collaborative inference.

Bertino and Oh propose a gradient-driven feature selector to identify important local features and iteratively perturb key features of the malicious client to propagate adversarial influence while suppressing benign features. The method is implemented in the VFLAIR framework with SplitNN protocol and evaluated on MNIST, CIFAR-10 and multimodal settings (UCF-101 for video- and audio-based action classification). Initial experimental results show that perturbing only 10% of input frame features from the malicious client reduces global model accuracy by 21.13%. These initial results expose a critical susceptibility of VFL to inference-time adversarial manipulation and demonstrate that feature-level attacks can dominate collaborative predictions even in realistic multimodal settings. A paper will be prepared and submitted for publication to a conference during the summer.

3.1.3 Benefits of use-inspired research

Research Impact. As of May 2026, ACTION is acknowledged in more than 300 manuscripts, most of which have been published in peer-reviewed venues. The remaining are under review at various conferences. The Institute's Google Scholar Profile shows an h-index of 44 [1].

Amazon Nova AI Challenge. After a head-to-head tournament designed to test the safety of AI coding models under adversarial conditions, Team PurpCorn-PLAN from the University of Illinois, Urbana-Champaign was named the winning defending team in the first-ever Amazon Nova AI Challenge [38]. PurpCorn-PLAN's roster includes ACTION researchers **Hadjer Benkraouda**, a PhD student at UIUC and the Institute's new Student Advisory Council co-lead, and **Gang Wang**, Professor of Computer Science at UIUC and the lead of ACTION's SEC-1 research thrust.

PurpCorn-PLAN's research, inspired by on-going work along ACTION's security research thrusts, has the potential to have a significant impact on various industries, including healthcare, finance, and transportation, where secure AI systems are critical for protecting sensitive information and preventing cyber threats.

DARPA AI Cyber Challenge (AIxCC). Team Shellphish, founded by ACTION Director **Giovanni Vigna**, finished in 5th place at the DARPA AI Cyber Challenge final, held at DEF CON 33 in Las Vegas in August. The technology driving Shellphish's automated cyber reasoning system drew heavily from ACTION research, demonstrating the institute's ability to develop an AI system that can detect and patch vulnerabilities without human assistance. The Shellphish CRS, dubbed "ARTIPHISHELL" reportedly had the best telemetry among the final seven teams and the highest patching success rate (>95%) [39].

Benchmarks for Use-Inspired Research. Over the past year, ACTION researchers produced six benchmarks spanning the Institute's research thrusts, from AI reasoning and agent behavior to cybersecurity and human-AI collaboration. Two benchmarks focus on LLM safety and reasoning quality in high-stakes contexts. *SOSBENCH* [40] evaluates whether LLMs maintain safety alignment when reasoning about scientific knowledge. This is critical for any organization deploying AI in cybersecurity. *GraphEval36K* [41], published at NAACL 2025, provides a large-scale dataset for assessing LLM performance on graph-structured reasoning tasks fundamental to network analysis and supply chain security. Both offer rigorous test beds for evaluating model reliability before enterprise deployment.

The *Visual Aesthetic Benchmark* [42] evaluates whether frontier models can make reliable judgments about visual quality with human-like nuance. An AI that can distinguish authentic from manipulated visual content has clear operational relevance in an era of synthetic media and AI-generated disinformation. Moreover, the benchmark's evaluation methodology is directly transferable to detecting AI-manufactured phishing content and social engineering artifacts.

Two benchmarks address the challenge of securing autonomous agents in physical and digital environments. *RoboJailBench* [28] systematically evaluates adversarial attacks and defenses against embodied robotic agents. Developed in direct coordination with GATE, which provides the cyber-physical infrastructure for deploying and stress-testing attack and defense agents against embodied AI systems, this benchmark gives a principled framework for evaluating autonomous systems before operational deployment. The *Yes-Man Syndrome* [29] addresses the tendency of AI agents to comply uncritically rather than abstain when appropriate. Although more subtle, this remains a failure mode in human-AI teaming scenarios. When working alongside security operations personnel, an agent that never declines can be as dangerous as one that is overtly adversarial.

Finally, *Beyond BeautifulSoup* [43] benchmarks LLM-powered agents on web scraping tasks that underlie a vast range of cybersecurity workflows, from threat intelligence gathering to vulnerability monitoring. It gives a principled basis for selecting and deploying such tools in real security operations.

Taken together, these six benchmarks embody ACTION's commitment to research that is simultaneously scientifically rigorous and operationally grounded. These benchmarks offer practical tools for procurement, red-teaming, and system evaluation, lowering the barrier between research innovation and real-world adoption.

3.2 Nurture and grow the next generation of talent

3.2.1 University and college education

Summer Research Experiences. ACTION hosted summer research experiences at Rutgers University and at UC Santa Barbara.

At Rutgers, in collaboration with a summer REU site at the Center for Discrete Mathematics and Theoretical Computer Science (DIMACS) [44], two students mingled

with 20 other students from DIMACS while working on projects of relevance to ACTION. The students **William Guo** from University of Pennsylvania and **Edward Xiong** from MIT joined **Professor Jie Gao** on a project that models learning in a social network. The group's project, meant to impact downstream ACTION research, for instance in groups of agents tasked with network defense, led to a paper that was accepted at the International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS'26).

UCSB hosted seven participants for their 8-week summer research experience. Students primarily came from the California State University system, where the mission of the university focuses on undergraduate teaching and practical workforce preparation. **Brian Lee** was from the University of Chicago, having learned of ACTION through his research advisor and **Professor Nick Feamster**. The other two students learned of ACTION through recruitment presentations at UCSB. Six of the seven participants at UCSB (and one high school student, see below) were mentored primarily by **Dr. Yigitcan Kaya**, a postdoctoral scholar at UCSB, and **Professor Christopher Kruegel**. Professor **João Hespanha** mentored the seventh UCSB participant, **Erik Feng**, who expressed interest in using adversarial AI for game theory research.

1. Megan Gross, Master's student, San Jose State University (California State University)
2. Anoushka Sawant, Master's student, San Jose State University (California State University)
3. Pedro Gonzalez, Bachelor's student, California State University, San Bernardino
4. Arjun Chopra, Bachelor's student, California Polytechnic University, San Luis Obispo (California State University)
5. Brian Lee, Bachelor's student, University of Chicago
6. Saqif Ayaan, University of California, Santa Barbara
7. Erik Feng, University of California, Santa Barbara

All participants at UCSB and Rutgers were screened through an online application process involving multiple rounds of review. ACTION management and the ACTION Student Advisory Council led the first two rounds of vetting, selecting students who had the best fit for the program and were positioned to receive the most value from participation. Faculty advisors at the various ACTION sites had the final say in onboarding participants.

During their summer research experiences, all participants were given opportunities to network with other summer undergraduate programs on their respective campuses and

to join seminars on professional development, presentation skills, and multidisciplinary research. UCSB and Rutgers also organized zoom sessions for summer research participants to meet and share their projects/results with each other. The online gathering proved to be a great success and we plan to continue or even expand such efforts in future summer programs.

Deliverables from ACTION's summer research program included abstracts, posters, and presentations. Appendix D summarizes the program and its outcomes.

Of note, the work led by **Megan Gross** and titled "COVERTLENS: Uncovering Concealed Harmful Behaviors in LLMs via Activation Analysis" appeared at the NeurIPS 2025 workshop. ACTION sent Ms. Gross to San Diego, CA to present her work in December 2025.

New graduates. This reporting period marks a significant point in the Institute's progress. A significant number of undergraduate and graduate students who have contributed more than two years to ACTION research have graduated.

- Kevin Lu, PhD in Mathematics from Rutgers University, May 2026
- Kaiyuan Zhang, PhD in Computer Science from Purdue University, 2026. Dr. Zhang will join Rutgers University as a faculty member in the ECE Department.
- Saurabh Sharma, PhD in Computer Science from UC Santa Barbara, May 2026. (co-advised by Jie Gao at Rutgers).
- Hadjer Benkraouda, PhD in Computer Science from UIUC, June 2026. (Expected)
- Hasan Amin, PhD in Computer Science from Purdue University, June 2026. (Expected)
- Madeline Blischke, PhD in Electrical and Computer Engineering from UC Santa Barbara, June 2026. (Expected)
- Doguhan Yeke, PhD in Computer Science from Purdue University, June 2026. (Expected)
- Saastha Vasan, PhD in Computer Science from UC Santa Barbara, July 2026. (Expected)
- Chengyuan Deng, PhD in Computer Science from Rutgers University, October 2026. (Expected)
- Prathamesh Dharangutte, PhD in Computer Science from Rutgers University, October 2026. (Expected)
- Vikrant Ashvinkumar, PhD in Computer Science from Rutgers University, October 2026. (Expected)

Project: zero2hero: Leveraging CTF-based Tools for Cybersecurity and AI Education

The national critical infrastructure is under constant attack from state-sponsored hackers who use sophisticated tools and techniques to gather intelligence, steal secrets, and disrupt services. Therefore, the training of the cybersecurity workforce has become a matter of national security, a problem that is exacerbated by the increasingly central role of artificial intelligence in devising security solutions.

To address this need, **Giovanni Vigna and Tonya Fields** created a new educational experience called "zero2hero," in which students learn cybersecurity and AI concepts without requiring any previous knowledge of these topics.

zero2hero is multifaceted, with deliverables developed at Norfolk State and UC Santa Barbara. At NSU, students explored and documented their experiences with two separate CTF-based tools: picoCTF [45] and a paid SANS subscription [46]. They created annotated walkthroughs that can supplement a student's learning via these online platforms. NSU students prepared early drafts of the documents over this reporting period and have plans to refine the drafts over the summer term. Once complete, the students at UCSB will be able to test the refined materials.

At UCSB, the deliverables took the form of two modular courses on vulnerability analysis. The first course is basic, where the goal is to describe security vulnerabilities at various levels of abstraction. Hardware vulnerabilities are a notable exception. They are out of scope.

The second course, focusing on vulnerability analysis and exploitation, is much more hands-on, extending the conceptual materials taught in class with specific CTF challenges and modules from pwn.college. zero2hero introduces 40 new CTF challenges developed exclusively for the course. Each challenge includes a walkthrough and AI prompts to create agents that support students in their learning.

Appendix D summarizes zero2hero's goals and achievements, including its two courses and 40 new CTF challenges. In the next reporting period, just as students at UCSB will test NSU's zero2hero materials, so too will NSU students test UCSB's courses.

Project: Composable Curriculum for Cybersecurity in Machine Learning and Generative AI: A Collaborative Educational Initiative

Radha Poovendran (U Washington), Wenke Lee (Georgia Tech), Bhaskar Ramasubramanian (Western Washington University) and Giovanni Vigna (UC Santa Barbara)

Cybersecurity is undergoing a rapid transformation due to the integration of machine learning (ML) and generative AI (GenAI) into components of the critical infrastructure. Threats now arise not only from traditional attack vectors, such as remote code execution and authentication bypasses, but also from model-specific vulnerabilities, such as adversarial examples, data poisoning, model leakage, and prompt injection. Existing curricula, often static and siloed, fail to reflect these evolving realities.

There is a critical need for a flexible, modular, and hands-on curriculum that covers these advanced concepts at the intersection of AI and Cybersecurity and promotes deep understanding through self-guided, inquiry-based learning supported by AI tutors.

This reporting period, ACTION's faculty leaders in education and outreach proposed and received a $2M award from NSF that will support expanded educational activities and provide nearly $1M in stipends for increased participation. Specifically, the grant outlines the development of new coursework and modules for an undergraduate and graduate curriculum. Proposed courses include:

- Machine Learning for Cybersecurity (ML4Cyber)
- AI Alignment for Safe Downstream Applications
- AI for Edge Intelligence (TinyML)
- AI/ML based Security Analytics
- Robustness of AI to Adversarial Manipulation
- Iterative Approaches to AI/ML Security

ACTION's goal is to develop these educational modules in time for the new NSF AI+Cybersecurity Summer Faculty Institutes. The institute also aims to impact programs like the National Applied AI Consortium, Level Up AI, and NAIRR Pilot Expansion, all of which were not included within the Institute's initial scope.

Deliverables this reporting period include a teacher/tutor chatbot by **Radha Poovendran** and a novel graduate level course by **Giovanni Vigna** titled *AVATAR:*

Autonomous Vulnerability Analysis, Triaging, And Repair. Appendix D summarizes these and the curriculum's other components.

In addition, **Wenke Lee, Radha Poovendran, and Bhaskar Ramasubramanian** have begun the adaptation of current classes to fit the proposed curricular models. Courses include

1. Introduction to Information Security, Network Security, and Infosec Labs
2. Machine Learning for Cybersecurity (ML4Cyber), including labs/rubrics for students to understand and do self guided evaluation
3. AI for Edge Intelligence (TinyML)

Project: AI+Cybersecurity MOOC

Dawn Song (UC Berkeley) and others as needed

Besides lectures, the proposed course will also include interactive lab assignments, hackathon, and a capstone project. We will also feature guest speakers from leaders in AI and cybersecurity across academia and industry. The course will culminate in a global AI & Security Hackathon, inviting participants to develop innovative tools and solutions.

The planned activity is over 1 year, with the first half year on designing and developing the course and the second half year on delivering the course. The course will cover key and emerging topics across most of the relevant research thrusts from ACTION. We also plan to integrate with the Globally Accessible Testing Environment (GATE) by designing labs and capstone projects that leverage GATE for hands-on security evaluation, adversarial testing, simulations, etc.

Through this new MOOC, we aim to educate and empower the next generation of technologists to build secure AI systems and use AI for securing the digital world. We envision a growing and engaged community of developers, students, and professionals working collaboratively at the frontier of AI and security.

The AI+Cybersecurity MOOC remains in development. We anticipate a course outline and timeline to be available for the Year 4 site visit.

3.2.2 Pre-college education

Summer Research Experience for High School Students. Professor Kruegel (UCSB) hosted a high school student for a summer research experience, as he did in 2023 and 2024. The interning student from San Jose, CA joined a group of summer research participants on one of the ACTION projects. He made regular contributions to the project, working directly with undergraduate colleagues and the postdoc mentor. Specifically, the high school participant was responsible for defining rules for the different tiers of redaction and contributing to the evaluative framework. The project, titled REDACTBENCH, developed a framework for using LLMs to redact confidential information from documents.

Summer CTFs Training Camps. In July, Norfolk State University hosted their first ACTION-sponsored Capture the Flag training camp targeting high school students. The week-long event brought seven high school teachers from Norfolk, Virginia Beach, and Newport News School Systems, and 25 high school students to NSU for participation in CTF exercises as a hands-on method for learning tenets of cyber security.

NSU will host these camps again in 2026. They will recruit the ACTION Evaluation team for assessment.

3.2.3 Workforce development

Summer CTF and Courses for Students and Veterans. Norfolk State University used their CTF training camps in July 2025 to successfully enroll five U.S. veterans, in addition to the 22 undergraduate students. These camps ran parallel to the high-school level courses described in Section 3.2.2. NSU offered the courses in the evening to accommodate participants' work schedules. **Tonya Fields** has arranged for NSU to host these camps again in 2026. She has already contacted the ACTION Evaluation team for help with program assessment, which is something she was unable to do for her 2025 courses.

3.3 Broadening participation and diversity

ACTION's approach to broadening participation is grounded in NSF's core principle that the benefits of scientific discovery should create opportunities for all Americans everywhere. The institute's strategy emphasizes expanding access based on geography, institutional type, socioeconomic background, and career stage. These are characteristics that have historically determined who does and does not have access to cutting-edge

research in AI and cybersecurity. All of ACTION's broadening participation activities are designed to reach communities and institutions where such opportunities have been scarce.

CTF-Based Outreach. Capture the Flag competitions have proven to be one of ACTION's most effective and scalable tools for broadening participation. As noted above in Section 3.2. Nurture and Grow the Next Generation of Talent, our Year 3 CTF outreach continued to grow in both reach and ambition. We include these efforts under Broadening Participation with notes that Norfolk State University's camps in July 2025 drew 25 high school students from school systems across the Hampton Roads region, including participants from Chesapeake, Portsmouth, Suffolk, and Norfolk. These are communities with limited prior exposure to research-grade cybersecurity training, demonstrating ACTION's geographic and socioeconomic reach can extend well beyond its founding campuses. A separate night camp at NSU served 22 students and five U.S. veterans, recognizing that military veterans represent an underutilized talent pool for the cybersecurity workforce. The iCTF 2026 "Battle of the Agents" competition expanded its reach to high school teams for the second consecutive year, with monetary prizes drawing participation from school districts that would not otherwise have the resources to compete at a national level. See Section 3.2.2. for more details.

Teaching the Teachers. Responding directly to the SVT's identification of "teaching the teachers" as a high-leverage opportunity, ACTION held its first instructor clinic in June 2025 at the conclusion of a year-long Faculty Outreach Grant awarded in Year 1. The clinic was attended by teachers in Santa Barbara County, who were the target group of the grant. The clinic equipped high school mathematics and computer science teachers with the knowledge and tools to launch their own student-led cybersecurity programs, a force-multiplier approach that reaches students in communities where ACTION does not have a direct physical presence.

Also of note, NSU's CTF camp was attended by seven teachers from the aforementioned school districts. By training educators who return to their home institutions, ACTION can extend its impact to students across a wide range of geographic and institutional settings without requiring those students to travel to an ACTION campus. NSU has committed to hosting these camps again in 2026 and will recruit ACTION's evaluation team for formal assessment, enabling the institute to measure the effectiveness of this approach in a principled way.

The NSF education supplement has deepened ACTION's engagement with the broader faculty development ecosystem. The supplement provided an introduction to the NSF

Faculty Development Summer Institute on AI in Cybersecurity Education, a fully-funded program hosted by George Washington University and the University of Hawai‘i at Mānoa running June through July 2026. Although late to the curriculum planning for the Summer Institute, ACTION has proposed the incorporation of CTF challenges from its Agentish platform into the later weeks of coursework, where cybersecurity teachers are exploring courses on their own. The supplement has also opened a new connection to NSA's Centers for Academic Excellence in Cybersecurity (CAE) program, specifically those programs that NSF is targeting for Cybersecurity + AI integration. With nearly 500 CAE-designated campuses nationwide, spanning community colleges, regional universities, and institutions in every state, this community represents one of the most significant geographic and institutional multipliers available to ACTION. The institute hopes to begin a formal collaboration in the next reporting period.

Undergraduate and Graduate Outreach. Summer research experiences at UCSB engaged students from the California State University system, a network of institutions whose mission centers on access and practical workforce preparation for students from a wide range of socioeconomic and geographic backgrounds. The program expanded to include a California community college student in summer 2026. One 2025 UCSB participant, Anoushka Sawant, went on to receive a fellowship with the NSF-funded National Community College Cybersecurity (NCyTE) Center, which prepares the next generation of cybersecurity educators. Her path is a concrete example of how ACTION's summer program connects students from a wide range of institutional backgrounds to national workforce development opportunities.

The Student Advisory Council also remains an active vehicle for broadening participation, ensuring that student voices shape the institute's outreach priorities and that the summer program continues to reach students who stand to benefit most.

The Year 2 SVT noted that broadening participation efforts were still somewhat concentrated on a few institutions and that ACTION should work to increase the breadth of outreach by involving a wider range of institutions and student organizations. In Year 3, the $2M NSF education supplement represents the most significant structural response to this concern, formally extending ACTION's educational activities to the University of Washington, Georgia Tech, and Western Washington University. The supplement has also opened connections to NSA's CAE program (497 campuses), CISA's CYBER.ORG platform, and the NSF Faculty Development Summer Institute, each of which represents a pathway to students and educators well beyond ACTION's founding institutions.

The SVT also flagged the proposed AI+Cybersecurity MOOC as a significant broadening participation opportunity, noting the potential to track engagement from populations beyond academia. The MOOC, led by Dawn Song at UC Berkeley, is currently in the planning phase, with course development scheduled for Year 4. ACTION's evaluation team has been notified of the MOOC and will attempt to develop outreach and tracking mechanisms to measure reach across institutional types, career stages, and geographic regions, consistent with NSF's current priorities for creating opportunities for all Americans everywhere.

ACTION is attentive to NSF's updated guidance on broadening participation, which emphasizes that all outreach activities must be open and available to all Americans and that engagement should be designed to expand access based on non-protected characteristics such as institutional type, geography, socioeconomic status, and career stage. All of ACTION's broadening participation activities are designed and implemented consistent with these principles.

3.4 Multidisciplinary integration and community building

Community building within ACTION accelerated in Year 3, driven primarily by the increased cross-campus collaboration required by the mini-project structure. Each of the 20 Year 3 mini-projects involved faculty from at least two separate ACTION campuses, producing a natural increase in researcher interaction. Appendix C lists the faculty on each project.

The Student Advisory Council (SAC) remains one of ACTION's most distinctive organizational strengths. Composed of PhD student representatives from across ACTION's partner institutions, the SAC serves multiple functions simultaneously: it provides the Director with a student perspective on institute governance, supports peer mentoring and seminar activities, and plays an active role in selecting summer research participants. This year, the SAC took on additional responsibilities, including the composition of a monthly newsletter highlighting research activities and opportunities for the broader graduate student community. SAC members also participated in a student exchange designed to encourage cross-campus collaboration and, in their capacity as the institute's intellectual scouts, began exploring the emerging topic of AI and satellite security as a potential future research direction. Graduate students who have passed through the SAC have described the relationships built across campuses as

one of ACTION's most tangible and lasting benefits, making conferences, workshops, and collaborative projects feel more welcoming and productive.

Managing Director Robinson participated in the 3rd Annual Summit for Digital Trust at EPFL, Switzerland, engaging with program managers and directors from peer institutions as well as representatives from philanthropic organizations. He will continue to cultivate these relationships during future reporting periods. See Section 3.7 for more details on this collaboration.

Finally, ACTION maintains informal contact with alumni of its summer research programs, tracking how participation has shaped participants' academic and career trajectories. While some have moved away from cybersecurity, others have continued in the field, a reminder that community membership, as the institute likes to say, does not end at graduation.

3.5 Multi-organizational synergies/achievements

ACTION's most significant multi-organizational achievement this reporting period was the receipt of a $2M NSF education supplement. Developed under UCSB's leadership and in close collaboration with the University of Washington and Georgia Tech, who will receive the majority of supplemental funds over the next three years, the supplement also incorporates Western Washington University as a new subawardee. The specific plans for the supplement are outlined in Section 3.2.1.

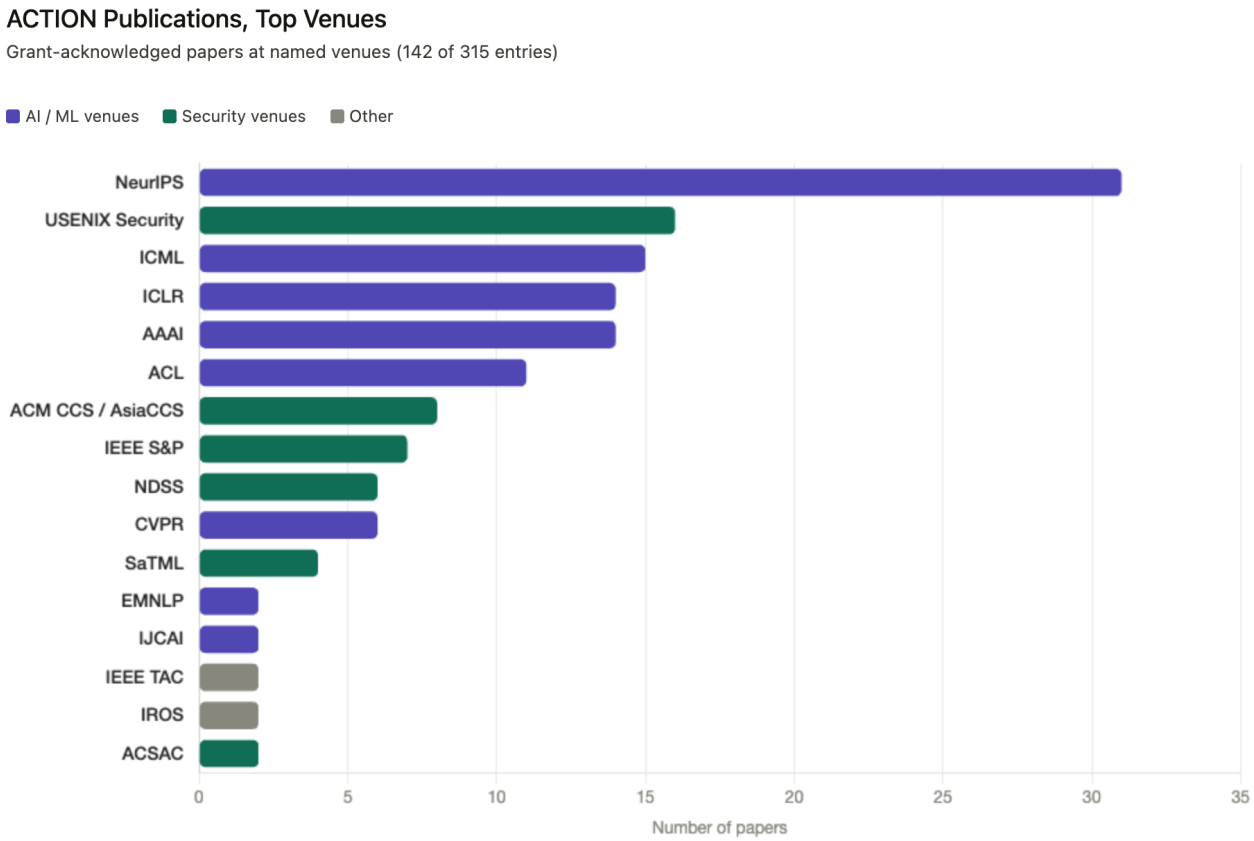
Additional multi-organizational achievements this year include several collaborative research deliverables spanning multiple ACTION campuses. Published benchmarks MABEL [31] and the Visual Aesthetic Benchmark (VAB) [42] resulted from cross-campus collaborations, as did CVE-GENIE, TOUCAN, and MAGPIE, all of which also engaged researchers from outside the ACTION community, including Arizona State University and Boston University. These projects are discussed in detail in Sections 3.1.2 and 3.6.

Institutionally, ACTION's mini-project funding model continued to enforce cross-campus collaboration as a structural requirement, ensuring that multi-organizational synergy is embedded in the research process rather than left to emerge organically. The Executive Committee now includes representatives from five institutions, and the SAC, described in detail in Section 3.4, continued to provide cross-campus student representation and leadership development across the institute.

3.6 Knowledge transfer

ACTION's research is generating meaningful traction well beyond the academic community, reaching federal agencies, industry partners, and the broader public through keynote presentations, tool releases, media coverage, and direct government engagement.

Research Dissemination. The graph below shows where we have shared our research results over the past three years. We colored the bars by venue type, with purple for AI/ML venues and teal for security venues. NeurIPS is by far our most common venue for dissemination, reflecting the strong foundational AI output of ACTION's researchers, especially Bo Li, Bing Liu, and Radha Poovendran's groups. USENIX Security ranks second among all venues, showing that ACTION's cybersecurity work is landing at the field's top venue. Finally, SaTML is a notable presence. This is a newer, specialized venue at the AI/security intersection.



Keynotes and Invited Presentations. Director **Giovanni Vigna** delivered the opening keynote at the 2025 ACM Conference on Computer and Communications Security (ACM CCS) in Taipei, Taiwan. His talk, titled "Autonomous Vulnerability Analysis, Triaging, and Repair: A Historical Perspective," took a decade-long view of automated vulnerability analysis from the DARPA Cyber Grand Challenge in 2016 through the AI Cyber Challenge in 2025, highlighting lessons learned and the transformative role of large language models in the emerging era of autonomous cyber reasoning systems.

Dr. Vigna also briefed the NITRD Cyber Security and Information Assurance Interagency Working Group (CSIA WG) at its July 2025 meeting, reporting on ACTION's progress to a body that coordinates federal cybersecurity R&D under the National Science and Technology Council. In addition, Dr. Vigna delivered a keynote at Amazon's 3rd Annual Bot Buster's Summit (August 2025), an internal conference bringing together experts from across Amazon retail and AWS focused on detecting and mitigating automation threats.

Professor **Wenke Lee** delivered the U.S. keynote at the NSF/CSA Singapore Workshop on Cybersecurity Innovations (May 7–8, 2026), arguing that AI agents are poised to develop most, if not all, software, and that security must be embedded throughout the development lifecycle rather than treated as an afterthought. Dr. Lee's work supporting Thrusts SEC-1 and SEC-2 and described in Section 3.1 and Appendix D of this document, helps justify his claim.

João Hespanha and **Bing Liu** delivered a substantial number of ACTION-related invited talks during this reporting period:

- **Hespanha, João**. Plenary talk, 15th IFAC Workshop on Adaptive and Learning Control Systems (ALCOS), Mexico City, July 2025.
- **Hespanha, João.** Invited talk, Cyber-Physical Systems Research Center, UC Santa Cruz, November 2025.
- **Hespanha, João.** Invited session talk and workshop, 64th Conference on Decision and Control, Rio de Janeiro, December 2025.
- **Hespanha, João.** Pathways to Science Panel, 64th Conference on Decision and Control, Rio de Janeiro, December 2025.
- **Hespanha, João.** Invited talk, Oden Institute for Computational Engineering and Sciences, University of Texas, Austin, January 2026.
- **Hespanha, João.** Invited talk, Information and Decision Science (IDS) Lab seminar series, Cornell University, April 2026.
- **Liu, Bing.** "Optimal Continual Learning for Autonomous Agents and Robots." Invited talk, NVIDIA, April 2025.

- **Liu, Bing**. "Autonomous Post-Deployment Continual Learning." Invited talk, ICCV-2025 Workshop on Multimodal Continual Learning, October 2025.
- **Liu, Bing.** "Is Continual Learning Easy with Foundation Models." Invited talk, ICML-2026 Workshop on Continual Adaptation at Scale, Seoul, South Korea, July 2026.

In September of the next reporting period, **Dongyan Xu** will deliver a keynote address to Cyber Sweden 2026 [47]. This keynote marks a continued exchange with Research Institutes of Sweden, which began in 2024 during our participation in the second Summit for Digital Trust at EPFL and continued with exchanges coordinated by DHS and Swedish Civil Contingencies Agency.

Tools, Datasets, and Industry Engagement. ACTION's open-source tools and datasets attracted significant industry attention this reporting period. TOUCAN, co-developed with IBM and released publicly on GitHub, was highlighted in an IBM blog post validating its quality and utility for the broader AI community [48]. CVE-GENIE drew immediate industry traction upon its arXiv release, generating outreach from CISCO, Intel, NIST, Red Hat, ChainGuard, and others, as well as invited presentations at Penn State, Brown, and UC Santa Barbara [49]. Full descriptions of both tools are provided in Section 3.1.2.

ACTION researchers at the University of Washington released *JobBench* [50], a new open-source evaluation standard for the real-world utility of AI agents in workplace settings. Rather than measuring what AI agents can do, *JobBench* measures what tasks human professionals want AI to absorb and then assesses how reliably models can perform them. Early results from the *JobBench* Leaderboard reveal that even frontier models frequently struggle with the sustained precision required for high-volume, repetitive tasks.

Of note, all tools and datasets referenced in this report and prior reports are publicly available. Most appear on arXiv at minimum, with many also published in peer-reviewed journals and conference proceedings; code and datasets are available on GitHub. We are currently curating links on the institute's website to further facilitate knowledge transfer.

Competition Achievements. ACTION researchers claimed two notable competition results this reporting period: a first-place finish and $250,000 prize at the Amazon Nova AI Challenge (Team PurpCorn-PLAN, UIUC) and a fifth-place finish at the DARPA AI Cyber Challenge final at DEF CON 33 (Team Shellphish), where the team recorded the

highest patching success rate among finalists. Full descriptions of both achievements are provided in Section 3.1.3 [38, 39].

Media Coverage. ACTION's research attracted national media attention this reporting period. **Giovanni Vigna** and **Dawn Song** were interviewed for an article in *The Atlantic* by Matteo Wong titled "Chatbots are Becoming Really, Really Good Criminals," [51] which described the AI hacking arms race and referenced both **Vigna's** research and **Song's** CyberGym experiments at UC Berkeley [52]. The article reached a broad general readership, translating ACTION's research agenda into terms accessible to non-specialist audiences.

Government Engagement. Science writers from the U.S. Government Accountability Office (GAO) interviewed **Director Vigna** as part of a GAO briefing on artificial intelligence. The discussions covered opportunities to improve proposed policy options and potential implications or concerns with their implementation. While the contents of the meeting remain internal to GAO at this time, they will be made available to Congress later in 2026.

We also meet monthly with scientists from DHS Science and Technology and CISA to share research achievements and provide early access to new manuscripts.

Academic Community Building. ACTION faculty play active roles in shaping the research venues and communities that serve the institute's agenda. **Berkay Celik** co-chairs the PRISM Workshop at NDSS 2026, a new workshop on Attack Provenance, Reasoning, and Investigation for Security in the Monitored Environment. **Christopher Kruegel** serves on PRISM's steering committee, with Palo Alto Networks sponsoring the Best Paper Award.

Veridical Data Science. ACTION's foundational AI research has always emphasized the importance of responsible, reliable, and reproducible learning and reasoning in cybersecurity. From the outset, the institute's AI-1 thrust identified **Dr. Bin Yu's** Predictability, Computability, and Stability (PCS) framework for Veridical Data Science as a principled foundation for end-to-end quality control in agent sensing-reasoning pipelines. While the cybersecurity-specific enhancements to PCS outlined in the institute's S&IPv1 and S&IPv2 were deprioritized as ACTION's AI-1 thrust evolved in Year 3, the institute continues to see value in the framework.

The framework is gaining traction in education. Dr. Yu joined a panel on "Teaching Veridical Data Science" at the 2025 Joint Statistical Meeting of the American Statistical Association, with the conversation subsequently published in the Journal of Statistics and Data Science Education. ACTION views incorporating PCS into its cybersecurity agent pipelines as a high-value direction for Year 4, particularly as ensuring truthful, auditable AI outputs becomes increasingly critical in adversarial security environments. Accordingly, we have added a thread for VDS to the S&IPv4.

Collaborative Project Management. As part of a collaboration with members of the Project Management Special Interest Group in AIVO, ACTION Managing Director **Tim Robinson** co-authored a study on project management practices across 25 NSF National AI Research Institutes, examining how project managers foster collective intelligence, improve operational efficiency, and cultivate cross-disciplinary relationships. The paper provides actionable recommendations for funding agencies, academic institutions, and research organizations, and represents a direct contribution to the science of running large-scale, interdisciplinary AI research programs.

TABLE 6 INTELLECTUAL PROPERTY

IP License Number or Name	IP License Title or Name	IP Category: FP, PP, C, T	Brief Description of Technology	Owner of IP	Year Awarded
none					

FP= full patent; PP = provisional patent; C= copyright; T= trademark

There is no reportable intellectual property. All research results have been publicly disseminated.

TABLE 7 - TECHNOLOGY TRANSFER

		Industrial Application		
Adopting Company or Partner	Technology	When transferred (date)	Use in Company	Impact (e.g., cost savings; productivity gain, financial and other market impact etc.)
Open Source Community; IBM (co-developer)	TOUCAN Dataset	July 2025	Co-developed with IBM and released publicly on GitHub. IBM highlighted TOUCAN in a company blog post, validating its quality and utility for the broader AI community. The dataset is freely available for any organization to use in training tool-using AI agents.	Largest publicly available dataset of its kind (1.5M data points from ~500 real-world MCPs). Establishes a new benchmark for tool-use training data quality and diversity, with IBM's co-development and endorsement lending significant credibility to its adoption by the open-source AI community.
Red Hat; Curl (testing collaborators); Security research community	CVE-GENIE	September 2025	Red Hat and Curl collaborated with ACTION researchers on testing CVE-GENIE's automated exploit reproduction pipeline against real-world software environments. Released publicly via ArXiv in September 2025; outreach received from CISCO, Intel, NIST, ChainGuard, and others exploring potential use cases.	Demonstrated ability to reproduce verifiable exploits for ~51% of 2024-2025 CVEs at $2.77/CVE average cost. Significant industry interest from major technology and security organizations suggests strong potential for adoption in enterprise vulnerability management and security benchmark generation workflows. Source code to be released publicly on GitHub following publication.

Open Source Community; DARPA / ARPA-H	ARTIPHISHELL Cyber Reasoning System (Shellphish)	August 2025	Released as open source in August 2025 alongside all six other AIxCC finalist CRSs. DARPA is actively applying open-source CRS technology — including ARTIPHISHELL — to open and closed-source repositories to identify and patch vulnerabilities in critical infrastructure software. The deployment tutorial has been successfully replicated by multiple external users.	ARTIPHISHELL's approach, 60+ autonomous AI agents for vulnerability detection and patching, demonstrated the best patching success rate (>95%) and best telemetry among all seven AIxCC finalists, establishing it as a leading open-source CRS for autonomous software security.

TABLE 8 - START-UP FIRMS

Name of Firm	Contact Information at Firm	Date Established	Name of Principal & Relationship to Institute (e.g. faculty, student, graduate, if any)	Funding status (SBIR, 1st round, positive tax income, etc.)	Technology	Market Impact or Societal Benefit (in terms of value added)
Virtue AI	Bo Li	2024	Faculty	Series A	AI security guardrail and risk assessment technology (Bo Li research)	Provides advanced risk assessment and guardrail solutions tailored to enterprise needs.

Artiphishell	Lukas Dresel	2025	Graduate Student	Seed round	AI-native application security platform (UCSB vulnerability research)	AI-native application security platform focusing on triage and false-positive reduction.

3.7 Impact of the Institute as a nexus point for collaborative efforts

After three years of operation, ACTION is increasingly functioning as a connector between the AI and cybersecurity communities. We bridge government agencies, industry partners, international research institutions, and the broader academic ecosystem.

Scholarly Impact. As of May 2026, ACTION's Google Scholar profile [1] lists over 300 manuscripts since late 2023, with more than 8,000 citations, an h-index of 44, and an i10-index of 147. We believe this impact can be leveraged to strengthen and justify future nexus-building activities and collaborations, providing an evidence-based foundation for demonstrating ACTION's influence on the broader AI and cybersecurity research community.

Government and Policy Nexus. ACTION's relationships with federal agencies have deepened significantly during this reporting period. Monthly meetings with DHS Science and Technology and CISA provide a direct channel for sharing research advances with the agencies responsible for protecting U.S. critical infrastructure, while early manuscript access ensures that policymakers are aware of emerging findings before they are fully published. Through **Director Vigna's** interviews with GAO in early 2026, ACTION's research may help inform GAO's forthcoming work for Congress on AI policy.

ACTION also served as the representative for all NSF National AI Research Institutes at the 2025 Special Competitive Studies Project (SCSP) AI+Expo in Washington, D.C. [53]. **Managing Director Robinson**, postdoctoral researcher Cheng Xin (Rutgers), and AIVO Managing Director Steve Brown shared the latest advances from ACTION and its peer

institutes with an audience of more than 15,000 government, industry, and academic leaders at an event featuring 275 speakers. NSF also highlighted other federally funded initiatives at the booth, including America's Seed Fund and the Regional Innovation Engines. While ACTION did not attend the 2026 SCSP AI+Expo, the institute was represented by AIVO, who secured at least one follow-on meeting with a Washington D.C.-area contractor about a potential collaboration.

Industry Nexus. Through its partnership with IBM, ACTION has been introduced to CoSAI [54], a cross-industry initiative working to advance the development of secure and trustworthy AI systems. ACTION plans to engage more actively with this community over the coming 12-month reporting period.

Virtue AI (**Bo Li, Dawn Song**) and Artiphishell (**Lukas Dresel, Christopher Kruegel, Giovanni Vigna**) are two companies emerging from ACTION's research. These projects demonstrate that the institute plays a part in the generation of commercially viable technology with real-world security applications.

International Nexus. As we did in 2024, ACTION participated in the annual Center for Digital Trust (C4DT) Summit, hosted at EPFL in Lausanne, Switzerland in August 2025. Joining leading research groups from around the world, ACTION participated in a working group spanning topics from K-12 education and CTF competitions to AI policy in Europe, the United States, Asia, and the Global South. We have contributed to a collaborative manuscript, currently under composition and led by researchers from Warwick University and the Turing Institute, that attempts to identify a complete AI stack for the purpose of helping nation states define or refine a national policy on AI adoption and control. We believe these activities position ACTION as a contributor to global AI governance conversations, not merely a domestic research program.

Workforce and Education Nexus. ACTION's 2025 summer intern Anoushka Sawant's subsequent path to becoming an NSF-funded cybersecurity educator through the NCyTE fellowship is a concrete demonstration of how ACTION's research pipeline connects to the national workforce development mission.

One area where ACTION is quickly becoming a nexus point for collaborative efforts is education and workforce development for cyber security. We have demonstrated an interest in education since our institute’s inception and have developed some innovative tools, including new courses and the Agentish Platform [55]. We have also launched student-run cybersecurity clubs in high schools near ACTION campuses and have expanded the iCTF to include high schools.

Our efforts have introduced us to the notion of Career Technical Education (CTE), which is an expanding model in U.S. education. In northern Santa Barbara County there is a newly-launched vocational track at a high school that offers college-level credit and industry certification in cybersecurity. ACTION is poised to collaborate with this program on curriculum-appropriate challenges for future versions of iCTF. Additionally, members of this program have offered to help ACTION with walkthroughs for select challenges. In the process, they will help us create new training materials.

While exploring college-prep course work with local high schools, we learned of a new College Board Advanced Placement (AP) course in Cybersecurity that will be offered at US high schools in 2026-2027. Schools near ACTION campuses have already expressed interest in participating in these new AP courses for their students. These courses will end with a standardized test meant to demonstrate proficiency in the field. Passing these AP tests leads to articulated credits at participating colleges and universities.

ACTION is well positioned to support schools adopting the new course. The College Board is developing AP Cybersecurity, with consulting guidance from CISA in the Department of Homeland Security, and CYBER.ORG [5], a CISA-funded education nonprofit, is designing curricular materials that help teachers prepare students for the exam. ACTION meets regularly with DHS personnel, primarily on research, and we recently reached out to CISA's education initiatives. We have begun to explore how ACTION's educational materials, such as zero2hero and our iCTF challenges, could complement CYBER.ORG's curriculum and CISA's Cyber Defense Education and Training (CDET), CISA Learning, and National Initiative for Cybersecurity Careers and Studies (NICCS) programs.

IBM also has a number of educational initiatives, including skills development programs in place at US colleges and universities, including HBCUs. ACTION has been meeting with IBM's Talent Development coordinator, exploring new opportunities for transition.

And finally, the awarding of supplemental education funds to ACTION in mid 2025 has introduced the institute to new educational initiatives at NSF and has catalyzed new meetings with NSA's Centers for Academic Excellence in Cybersecurity (CAE). The CAE represents a community of 497 campuses with focused vocational training in Cybersecurity. We believe AI will eventually permeate each of the CAE programs, and look forward to transitioning our early work in training and education for agentic programming to each of these schools.

All of the above reflects a serendipitous realization of where ACTION sits at the intersection of government defense agencies (e.g. NSA, DOD, DHS), the NSF, and industry (via IBM). All of these organizations have separate agendas for workforce development. ACTION has an opportunity to connect them all, thereby strengthening the U.S.A.'s cyber defense workforce. We will continue to explore the potential opportunities in Year 4.

Further Impact. Two recent announcements highlight the growing influence of ACTION's faculty beyond the institute itself. ACTION faculty member Elisa Bertino was elected President of the Association for Computing Machinery (ACM), beginning a two-year term on July 1, 2026 [56]. Dr. Bertino's contributions to SEC-2, SEC-4, and cross-institute research on cellular network security and agentic memory are described throughout this report. Also, ACTION faculty member Ninghui Li, a contributor to Thrust AI-3 and SEC-2, joined Perplexity AI as the inaugural director of their new Secure Intelligence Institute [57]. Both appointments reflect the caliber of researchers ACTION has assembled and open concrete pathways for nexus building, with the broader computing research community through ACM, and with an influential AI industry player through Perplexity, that we expect to develop further in Year 4.

4.0 Evaluation and Assessment

ACTION's evaluation framework is grounded in a program logic model that maps the institute's activities to measurable outputs and intended short- and long-term outcomes across four programmatic areas: Outreach and Mentorship, Academic Preparation, Research Preparation, and Professional Development and Networking. Section 4.1 below provides an executive summary of the Year 3 evaluation report, drafted by the External Evaluation Team of Dr. Natalie Jones and Dr. Tarek Azzam. Section 4.2 includes a response from ACTION Management.

4.1 Evaluation Report | Executive Summary

Prepared by The UCSB Center for Evaluation & Assessment

Overview of Evaluation Activities

This past year, the evaluation team significantly deepened both the scope of its work and its engagement with the ACTION Institute. Core evaluation activities included the following:

Evaluation Tools and Data Collection

- Gathered feedback from key program stakeholders to update and refine the program logic model, ensuring evaluation efforts remain aligned with program goals and activities.
- Conducted in-depth interviews with students who participated in the ACTION Summer Internship Program, capturing qualitative data on participant experiences, outcomes, and program strengths and areas for growth.
- Updated and revised the iCTF student survey instrument; collected, analyzed, and synthesized both qualitative and quantitative data; and produced and shared findings briefly with program leadership.

Stakeholder Engagement and Partnership

In addition to core data collection activities, the evaluation team substantially increased the frequency and depth of its engagement with the program this year, including:

- Regular meetings with the program manager to maintain alignment and provide ongoing evaluation support.
- Attendance at monthly faculty meetings to stay current with program developments and build collaborative relationships with program stakeholders.
- Data sharing and data sensemaking sessions to support evidence-informed decision-making within the program, including findings sessions with the program manager after analysis of collected data.
- Collaborative input on survey design and evaluation protocols to ensure tools are rigorous, relevant, and responsive to program needs.

Summary of Key Findings

Prior Evaluation Findings

During the Winter quarter of 2025,49 respondents who participated in the iCTF in November 2024 completed a post-competition survey. Self-reported experiences were broadly positive across learning, collaboration, and satisfaction dimensions, with most respondents rating overall quality and engagement as Excellent or Good and the large

majority indicating they would recommend the competition to others. Collaboration and problem-solving items received the strongest agreement, while career and professional goal outcomes were present but more moderate. Open-ended responses were predominantly positive, with enjoyment, team-based learning, and AI and LLM challenge content cited most frequently as strengths. Recurring constructive themes included requests for greater challenge variety, more consistent logistical communication, and post-competition writeups. The survey instrument has since been updated, and direct comparisons of current-year findings should be made with that in mind.

ACTION Summer 2025 Internship Program

In-depth interviews with Summer Internship participants (n=5) yielded consistently positive feedback across all respondents. Key findings include:

- Participants unanimously described the program as a positive experience and a meaningful tool for both academic and career development.
- Mentorship was the most frequently cited driver of participant growth. Interns highlighted mentor availability, guidance, and encouragement as the most impactful element of the program.
- The program's focus on the intersection of artificial intelligence and cybersecurity was a strong draw for participants, and engagement deepened interns' interest in specializing in cybersecurity and exploring related career pathways.
- Participants reported meaningful gains in career clarity and academic direction, with many leaving the program with a clearer sense of whether to pursue graduate school or industry roles.
- The self-guided structure of the internship was broadly appreciated, though participants noted it may not be equally well-suited to all learning styles and needs.

Areas most commonly identified for improvement were structural in nature:

- Interns expressed interest in more organized collaborative and social events to build community across the cohort.
- Overall, participants reported that research skills and professional mindsets developed through the program were seen as transferable and valuable, regardless of whether interns planned to continue in research.

iCTF March 2026

Survey respondents (*n*=27) reported broadly positive experiences with the March 2026 iCTF competition. Key findings include:

Overall Competition Ratings (5-point scale) Educational: 4.42 | Worthwhile: 4.42 | Fun: 4.19
All 26 respondents reported the competition met or exceeded their expectations.

- Practical skill application was the most consistently self-reported outcome. Respondents indicated they were better able to apply practical technical skills (mean: 4.11/5) and collaborate effectively in teams (mean: 3.89/5).
- Retrospective pre/post ratings indicate perceived gains across all three AI and cybersecurity knowledge areas surveyed, with the largest self-assessed gain in awareness of how AI is used to automate real-world cybersecurity tasks (average change of +0.67).
- Career interest in cybersecurity specifically was the most elevated outcome in the career interest domain, with a mean rating of 4.33 out of 5.
- The most consistent constructive theme in open-ended responses was the volume and technical difficulty of agentic challenges, which a notable share of participants identified as a factor in their experience.

Proposed Evaluation Activities: Year Ahead

Building on the foundation established this year, the evaluation team proposes the following activities for the coming academic year:

- **Continued tool development and data refinement:** The evaluation team will continue developing and refining evaluation instruments and measures to ensure the program's data needs are met and progress toward program goals is tracked.
- **Sustained stakeholder engagement:** The team will maintain and deepen engagement with program leadership, the program manager, and faculty — including regular meetings, attendance at faculty meetings, and collaborative data sensemaking sessions.
- **Pre/post survey development for the Summer Internship Program:** The evaluation team will develop a pre/post survey instrument for summer interns informed by previous qualitative findings to capture both quantitative and qualitative data on participant outcomes, enabling more systematic tracking of growth across the internship experience.

- **Expanded support for faculty data collection:** As evaluation capacity grows, the team proposes expanding services to assist faculty with survey development and data collection support as needed, strengthening the overall evidence infrastructure for the program.

4.2 ACTION Response to Evaluation Report

We thank Dr. Jones for leading this year's evaluations and assessments, and Dr. Azzam for his guidance and feedback. We appreciate their willingness to work with us at a granular level to shape surveys and data collection to better capture the information we think will help us improve our activities and increase impact.

The monthly check-ins and attendance at faculty meetings have been especially helpful, allowing all of us to stay on the same page. We note the small sample size for the 2025 summer research experience. Hopefully, with a head start on pre-surveys, we can collect feedback from the participants beyond UCSB, including those on the semester system.

Also, we acknowledge the results of the 2025-2026 iCTF. This competition took an extensive amount of work from all of the researchers involved, despite being delayed by three months. Now that we have a benchmark for an agentic CTF, we expect a more accurate assessment of our impact. For example, at the next iCTF we will be able to prepare for security concerns at US high schools. Accessing the iCTF servers, with their dynamic challenges, proved difficult for many students trying to participate.

Finally, we agree with and look forward to the proposed evaluation activities in the year ahead. However, we would like to draw attention to the SVT's notes about program evaluation in our Year 2 Site Visit Report. The SVT has noted a number of weaknesses that center around an inability to measure the Institute's progress. As your evaluations in Year 3 concerned the education activities, we would use this response as a reflection on the logic model included in your formal report, and our successful progress against this model. While we have discussed many of the stated outputs and outcomes in greater detail above in this report, they deserve repeating in the context of evaluation.

Activity 1: Outreach and Mentorship

- Outputs achieved
    - NSU hosted its first ACTION-sponsored CTF training camp, reaching 25 high school students and seven teachers.

    - A separate NSU night camp served 22 students and five U.S. veterans.
    - iCTF 2026 "Battle of the Agents" expanded its high school division for the second consecutive year, attracting teams from diverse school districts nationwide.
    - Seven students participated in the 2025 UCSB summer research experience; two participated through the DIMACS REU at Rutgers.
    - Twice as many students will be supported in 2026 summer research experience, with host campuses being UCSB, UW, GeorgiaTech, and Rutgers. Host campuses have doubled each summer.
- Short-term outcomes:
    - iCTF survey respondents (n=27) reported career interest in cybersecurity at a mean of 4.33/5.
    - All 26 respondents who answered the expectations item reported the competition met or exceeded their expectations.
    - Internship participants (n=5) unanimously described the program as a positive and meaningful experience for academic and career development, with mentorship cited as the most impactful element.
- Long-term outcomes — early indicators:
    - Anoushka Sawant's transition from ACTION summer intern to NSF-funded cybersecurity educator through the NCyTE fellowship represents a concrete pipeline outcome.
    - The iCTF's second consecutive expansion to the high school division reflects growing institutional participation over time.

Activity 2: Academic Preparation

- **Outputs achieved:**
    - zero2hero, developed jointly by UCSB and NSU, introduced 40 new CTF challenges with AI-powered tutors for students with no prior cybersecurity knowledge.
    - The $2M NSF education supplement formalized new course development across UCSB, UW, Georgia Tech, and WWU, spanning six proposed courses at the undergraduate, graduate, and professional levels.

    - Existing courses at Georgia Tech (Introduction to Information Security, ~1,000 students/semester; Network Security; Information Security Labs), UW, and WWU are being adapted to incorporate AI+cybersecurity content.
    - AVATAR (Autonomous Vulnerability Analysis, Triaging, And Repair), a novel graduate course, was developed as an early supplement deliverable.
    - Cross-campus Zoom sessions were organized for summer research participants at UCSB and Rutgers in Summer 2025, providing virtual seminar experiences. Sessions across all four campuses supporting students in Summer 2026 are planned and ready for implementation.
- **Short-term outcomes:**
    - iCTF retrospective pre/post ratings showed gains across all three AI and cybersecurity knowledge areas, with the largest self-assessed gain in awareness of how AI automates real-world cybersecurity tasks (+0.67 average change).
    - zero2hero's AI-powered tutoring model represents a scalable approach to increasing student enrollment in foundational AI and security content.
    - The AI+Cybersecurity MOOC is in the planning phase, expected to reach a global audience in Year 4.
- **Long-term outcomes — early indicators:**
    - The education supplement's connection to the NSF Faculty Development Summer Institute and NSA's Centers for Academic Excellence in Cybersecurity (497 campuses) opens pathways for broad curriculum adoption well beyond ACTION's founding institutions.
    - ACTION is currently working to disseminate agentic CTF challenges and the Agentish platform widely, for education purposes.

Activity 3: Research Preparation

- **Outputs achieved:**

    - Twenty mini-projects spanning all eight research thrusts and GATE involved faculty from 12 institutions, representing a high level of cross-institutional collaboration.
    - Nine students and two postdoctoral researchers participated in summer research experiences at UCSB and Rutgers.
    - Summer intern Megan Gross presented "COVERTLENS: Uncovering Concealed Harmful Behaviors in LLMs via Activation Analysis" at the NeurIPS 2025 workshop. Students William Guo and Edward Xiong, working with Professor Jie Gao through the DIMACS REU at Rutgers, co-authored a paper accepted at AAMAS 2026. This is a direct pipeline from undergraduate research experience to peer-reviewed presentation.
    - As documented in the Strategic Implementation Plan (S&IPv3), the postdoctoral exchange program has been expanded and renamed the Postdoc/Senior PhD Scholar Exchange, opening participation to advanced PhD students.
    - ACTION's Google Scholar profile now lists over 300 grant-acknowledged manuscripts, an h-index of 44, and an i10-index of 147.
    - Multiple PhD students graduated this reporting period after two or more years of ACTION contributions.
- **Short-term outcomes:**
    - Internship participants reported meaningful gains in research skills and professional mindsets, describing these as transferable and valuable regardless of future career path.
    - Cross-institutional collaboration increased through the mini-project structure, which requires faculty from separate campuses to work together on defined deliverables.
- **Long-term outcomes — early indicators:**
    - Kaiyuan Zhang joining Rutgers University as a faculty member in the ECE Department represents a direct pipeline from ACTION graduate student to faculty researcher — precisely the long-term research preparation outcome the logic model envisions.
    - The growing publication record and h-index of 44 reflect a research program whose outputs are being built upon by the broader scholarly community.

Activity 4: Professional Development and Networking

- **Outputs achieved:**
  - The Student Advisory Council provided leadership development opportunities for student members, who played an active role in vetting and selecting summer research participants.
  - Summer program participants at UCSB and Rutgers attended professional development seminars on presentation skills, multidisciplinary research, and career pathways.
  - The NSU night camp served five U.S. veterans, extending professional development opportunities to a population with strong potential for cybersecurity careers.
  - Two startup companies founded by ACTION researchers, Virtue AI (Bo Li) and Artiphishell (Lukas Dresel), represent entrepreneurship pathway outcomes directly traceable to ACTION research and mentorship.
  - Potential to participate in CoSAI activities and an increasing network of industry collaborators point to opportunities for an expanding institute.
- **Short-term outcomes:**
  - Summer participants reported meaningful gains in career clarity, with many leaving the program with a clearer sense of whether to pursue graduate school or industry roles.
  - The iCTF's career interest in cybersecurity outcome (mean: 4.33/5) suggests the competition is effectively sparking professional interest among participants.
- **Long-term outcomes — early indicators:**
  - The Artiphishell startup, founded by ACTION graduate Lukas Dresel, represents a direct pipeline from ACTION research to cybersecurity entrepreneurship, one of the logic model's intended long-term outcomes.
  - Ongoing connections with IBM's Talent Development team and IBM's SkillsBuild program represent emerging pathways for increasing graduate job placement in cybersecurity fields.

We acknowledge that this current evaluation infrastructure, while deepening, is most fully developed on the education side of the program, to outreach and student development. Extending rigorous evaluation to research outputs, knowledge transfer, and industry engagement remains a Year 4 priority. The Google Scholar profile, h-index tracking, and publication counts provide a foundation for this sort of research output evaluation, and the technology transfer table developed this year offers a structured framework for tracking industry and government engagement over time. We will work with Dr. Jones and Dr. Azzam to develop evaluation instruments and data collection protocols that capture these dimensions more systematically in the coming year. We will begin this process through the further development of the current logic model.

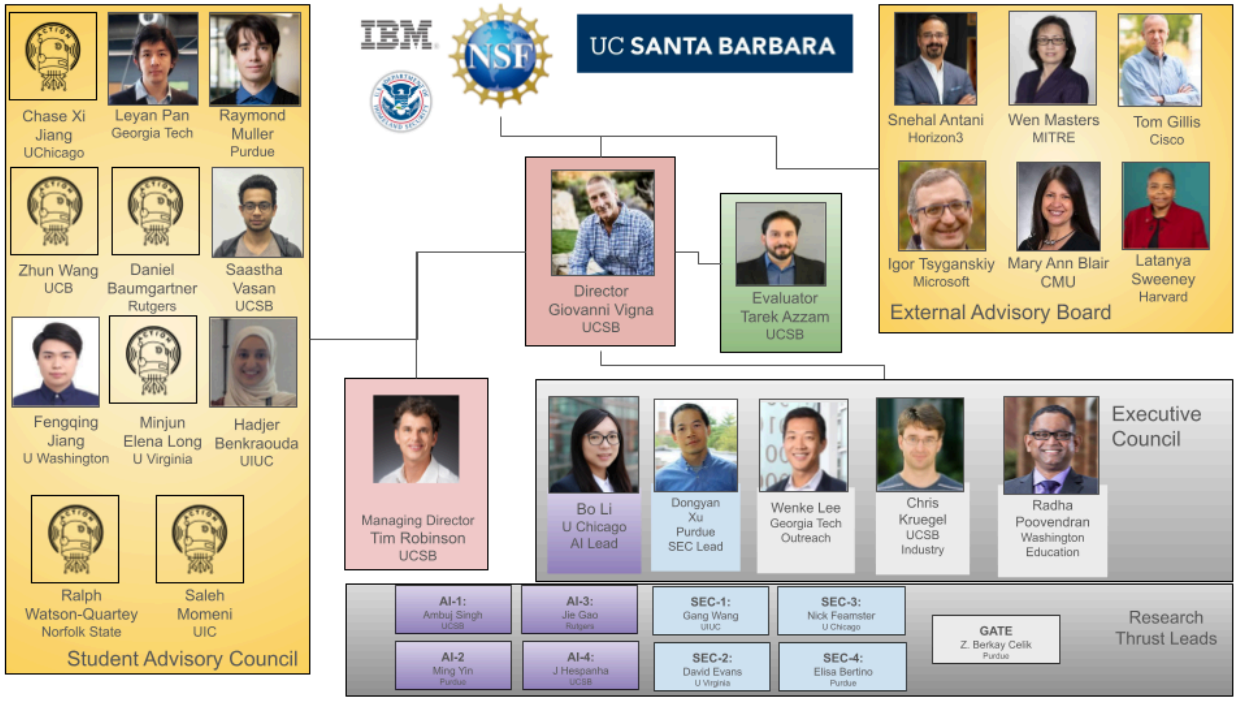
We also note that the logic model footnote commits to disaggregating data by Race and Ethnicity, Gender, and SES when possible. Consistent with NSF's updated guidance on broadening participation, ACTION will ensure that any such disaggregation is conducted in a manner that is open, non-exclusionary, and focused on creating opportunities for all Americans and will work with the evaluation team to design data collection instruments accordingly.

5.0 Infrastructure and Management

5.1. Configuration and Leadership

There have been no changes to ACTION leadership during this reporting period.

5.2. Organizational Chart



Although there have been no changes to the Institute leadership, we rearranged the above organizational chart to reflect the importance of Education, Outreach, and Industry partnerships. The rearrangement is meant to address the risks identified by the SVT comments. Namely, our prior organizational chart appeared to create a hierarchy of research results over educational impact. By listing the entirety of the EC at the same level, we hope to better exemplify our distributed priorities.

Individual Roles and Responsibilities

Giovanni Vigna, Director. The Director is responsible for all aspects of the Institute, including management, staffing, resource allocation, and communication with and reporting to NSF. The Director serves as the primary point of contact for the External Evaluator, the External Advisory Board, the Student Advisory Council, and primary campus (UCSB) leadership.

Timothy Robinson, Managing Director. The Managing Director is responsible for the daily management of the Institute and the coordination among its members, the

organization of events and meetings, the recruiting of talent, the administration of funds, and the interaction with entities external to the Institute.

Various, External Advisory Board. The EAB is composed of nationally and internationally recognized scientists and experienced industry leaders. Its mission is to provide guidance, advice, and oversight for all Institute activities, reviewing progress toward goals including research advancement, broadening participation, workforce education, and societal and technological impacts. The EAB provides internal advice and public accountability.

Various, Student Advisory Council. The SAC is composed of student representatives from the various ACTION institutions and provides the Director with the perspective of students and postdocs involved in the research. The SAC promotes student initiatives, supports student mentoring and seminar activities, and helps with the selection of ACTION interns. Each ACTION campus receives funding to support a graduate student representative annually.

Various, Executive Committee. The EC is the primary advisory body to the Director, advising on matters related to the integration of activities across institutions and research thrusts, progress in research projects, and the allocation of resources.

> **Bo Li, Foundational AI Research Lead.** Bo Li oversees the Institute's foundational AI research agenda, ensuring that the four AI research thrusts (AI-1 through AI-4) advance novel fundamental primitives for reasoning, learning, human-agent interaction, multi-agent collaboration, and strategic gaming. She coordinates with the Use-Inspired Research Lead and serves as a go-to resource for all activity related to foundational AI across the Institute.
>
> **Dongyan Xu, Use-Inspired Research Lead.** Dongyan Xu oversees the Institute's use-inspired cybersecurity research agenda, ensuring that the four SEC thrusts (SEC-1 through SEC-4) translate foundational AI advances into novel approaches for vulnerability assessment, threat detection, cyber threat intelligence, and response and recovery. He coordinates continuously with the Foundational AI Research Lead to maintain alignment between theoretical and applied research.
>
> **Wenke Lee, Broadening Participation and Outreach Lead.** Wenke Lee advises on all activities related to broadening participation, diversity, and outreach beyond the Institute's core organizations. His responsibilities include overseeing K-12 engagement, managing relationships with community organizations, and

ensuring that ACTION's outreach activities reach appropriate communities and institutions. He works closely with the Education Lead to align outreach with educational programming.

Radha Poovendran, Education Lead. Radha Poovendran oversees all of the Institute's educational efforts, including undergraduate and graduate curriculum development, MOOCs, workforce development programs, and faculty development activities. He works closely with the Outreach Lead and ensures that ACTION's educational outputs are aligned with the Institute's goals and responsive to stakeholder needs.

Christopher Kruegel, Industry Liaisons Lead Christopher Kruegel coordinates knowledge transfer to industry and serves as the primary liaison between ACTION researchers, industry members, and technology transfer professionals at the various institutions. His responsibilities include facilitating the transition of research tools and datasets into commercial or operational use, and cultivating relationships with companies and organizations that can advance or adopt ACTION's research outputs.

Individual Thrust Leads Thrust leads are senior faculty researchers responsible for directing research activities within their assigned thrust, coordinating collaborations across institutions, overseeing mini-projects, and ensuring that progress is reported and aligned with the Institute's goals. They report to the relevant EC pillar lead and work to foster the virtuous cycle between foundational and use-inspired research.

Tarek Azzam and Natalie Jones, Institute Evaluators. The Institute Evaluator monitors adherence to the Institute's broadening participation and management plans, conducts regular interviews with Institute personnel and participants, and ensures consistency between what is planned, what is reported by the Executive Committee, and what is perceived by participants.

5.3. Management Systems

Roles of advisory boards. The roles and responsibilities of the two ACTION advisory boards – External Advisory Board and the Student Advisory Council – have not changed during this reporting period.

New student representatives from various ACTION campuses have joined the SAC to replace those who have graduated in June 2025, or who stepped down from their roles. New members include **Zhun Wang** (UC Berkeley), **Minjun Long** (U Virginia), **Doguhan Yeke** (Purdue) and **Xi Jiang** (U Chicago). Also, **Saleh Momeni** joined the SAC as ACTION's first delegate from U Illinois, Chicago, and **Ralph Watson-Quartey** now represents Norfolk State University.

Graduate students from UIUC (**Hadjer Benkraouda**) and UCSB (**Saastha Vasan**) agreed to serve this year as co-chairs of the SAC.

Project Management. There have been no changes to project management during this reporting period.

Conflict of Interest Reporting. We do not have any current conflicts of interest.

5.4. Ethics

ACTION's research operates at the intersection of offensive and defensive cybersecurity, and the institute takes its ethical responsibilities in this space seriously. Several Year 3 research outputs, including CVE-GENIE, which automates exploit reproduction, and ARTIPHISHELL, an autonomous cyber reasoning system, represent dual-use technologies with clear potential for misuse. In each case, release decisions were made deliberately, with attention to responsible disclosure norms and the defensive value of making these tools available to the security research community. ACTION researchers follow the publication and disclosure standards of their respective venues, which for security research typically include coordinated disclosure practices and community review.

For agentic AI systems deployed in or connected to real infrastructure, ACTION researchers apply standard security research ethics principles: tools are evaluated in controlled environments before release, and datasets such as MABEL and TOUCAN are constructed from publicly available or synthetically generated data to avoid privacy implications.

The evaluation activities described in Section 4.0, including surveys and interviews with summer research participants and competition participants, are conducted under appropriate human subjects review protocols at the relevant institutions.

Finally, ACTION leadership is exploring supplementary ethics training options for institute participants. We evaluated a Coursera course in AI ethics developed by the Polytechnic University of Milan, recommended by the Special Competitive Studies Project, and will make a decision about its incorporation into institute requirements during Year 4. More broadly, ACTION intends to develop ethics content appropriate to the AI and cybersecurity intersection, an area not well served by general AI ethics curricula, and will report on progress in the next annual report.

References

[1] ACTION AI Institute, "Google Scholar Citations Profile," *Google Scholar*. https://scholar.google.com/citations?hl=en&user=KMn7qeAAAAAJ (accessed Jun. 1, 2026).

[2] H. Amin, M. Yin, and R. Khanna, "Align When They Want, Complement When They Need! Human-Centered Ensembles for Adaptive Human-AI Collaboration," in *Proc. AAAI Conf. Artif. Intell.*, vol. 40, no. 21, 2026.

[3] H. Yizhou Tian, H. Amin, and M. Yin, "Understanding the Effects of AI-Assisted Critical Thinking on Human-AI Decision Making," in *Proc. 2026 CHI Conf. Human Factors Comput. Syst.*, 2026.

[4] IBM, "Free Skills-Based Learning From Technology Experts," *IBM SkillsBuild*. https://skillsbuild.org/ (accessed May 31, 2026).

[5] CYBER.ORG, "Cybersecurity Education and Workforce Development," *CYBER.ORG*. https://cyber.org/ (accessed Jun. 1, 2026).

[6] Y. Zhang, K. Lu, Y. Zhang, J. Gao, L. Xia, and F.-Y. Yu, "Mixture of Complementary Agents for Robust LLM Ensemble," presented at the *NeurIPS 2025 Workshop on Socially Responsible and Trustworthy Foundation Models (ResponsibleFM)*, Dec. 2025.

[7] S. Sharma and A. Singh, "Emergence of Phase Transitions in Complex Contagions," *arXiv preprint arXiv:2603.18380*, 2026.

[8] C. Sardina, E. Polat, and A. Singh, "BottleneckMLP: Graph Explanation via Implicit Information Bottleneck," *OpenReview*. [Online]. Available: https://openreview.net/pdf?id=4qj7qO1fTJ

[9] E. Polat and A. Singh, "GLoSS: Natural Language Counterfactuals for GNNs via LLM Latent Steering," submitted for publication, 2026.

[10] C. Sardina and A. Singh, "Representations + Lie Algebra = Reasoning," submitted for publication, 2026.

[11] W. Guo, E. Xiong, and J. Gao, "Robust Sequential Learning in Random Order Networks," in *Proc. 25th Int. Conf. Auton. Agents Multiagent Syst. (AAMAS)*, Paphos, Cyprus, May 2026, oral presentation.

[12] C. Deng, J. Gao, S. He, and C. Wang, "Efficient Testing for Correlation Clustering: Improved Algorithms and Optimal Bounds," in *Proc. 14th Int. Conf. Learn. Represent. (ICLR)*, Rio de Janeiro, Brazil, Apr. 2026.

[13] B. Liu, X. Huo, and W. Lee, "Unlearning in Post-Deployment Learning Systems," submitted for publication, 2026.

[14] B. Liu, X. Huo, and W. Lee, "Beyond Fixed Models: Toward Living Systems That Learn on the Job," in preparation, 2026.

[15] B. Liu, X. Huo, and W. Lee, "Unlearning in Deployed Evolving Learning Systems," in preparation, 2026.

[16] M. Tan, M. A. Merrill, Z. Gottesman, T. Althoff, D. Evans, and T. Hartvigsen, "Inferring Events from Time Series Using Language Models," in *Proc. 64th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, San Diego, CA, USA, Jul. 2026. [Online]. Available: https://arxiv.org/abs/2503.14190

[17] M. Yin, "AI-Assisted Hidden Profile Decision-Making," poster to be presented at the *12th Int. Conf. Comput. Social Sci. (IC2S2)*, Burlington, VT, USA, Jul. 2026.

[18] A. K. Musaffar, A. Gokhale, S. Zeng, R. Tadayontahmasebi, X. Yan, A. K. Singh, and F. Bullo, "Learning to Lie: Adversarial Attacks on Human-AI Teams and LLMs," in *Proc. 14th Int. Conf. Learn. Represent. (ICLR)*, Rio de Janeiro, Brazil, Apr. 2026.

[19] A. K. Musaffar, A. Singh, and F. Bullo, "Detection of Adversarial Intent in Human-AI Teams Using LLMs," presented at the *ICLR 2026 Workshop on From Human Cognition to AI Reasoning: Models, Methods, and Applications*, Rio de Janeiro, Brazil, Apr. 2026.

[20] S. Zeng, M. Yin, X. Yan, F. Bullo, and A. Singh, "Investigating Language Models for Supporting Complex Group Decisions," in preparation, 2026.

[21] K. Zhang, X. Su, P.-Y. Chen, E. Bertino, X. Zhang, and N. Li, "Security Principles for Agentic AI," *Commun. ACM*, revision under review, 2026.

[22] S. Haddadan, C. Xin, and J. Gao, "Maximizing Coalitional Rank Improvement in a Social Network," submitted for publication, 2026.

[23] A. J. Li et al., "Green Shielding: A User-Centric Approach Towards Trustworthy AI," *arXiv preprint arXiv:2604.24700*, 2026.

[24] A. Bhowmik, S. Bopardikar, and J. P. Hespanha, "Data-Driven Robust Control Under Input-Output Stealthy Attacks," *IEEE Control Syst. Lett.*, vol. 9, pp. 2885–2890, Dec. 2025, presented at the *2026 Amer. Control Conf.*, May 2026.

[25] K. G. Vamvoudakis, J. I. Poveda, S. Anderson, J. P. Hespanha, M. Krstic, I. R. Petersen, M. Tang, A. Wadi, and Q. Zhu, "A Tutorial for Foundations in AI and Quantum Systems: From Control and Game-Theory to Stochastic Deception," Univ. California, Santa Barbara, CA, USA, Tech. Rep.

[26] M. Blischke and J. P. Hespanha, "Optimal Control of Entity-Based Systems via Koopman Representations with Product Density Observables," *IEEE Open J. Control Syst.*, 2026.

[27] S. Anderson and J. P. Hespanha, "Learning with Contextual Information in Non-Stationary Environments," in *Proc. 7th Annu. Learn. Dyn. Control Conf. (L4DC)*, Ann Arbor, MI, USA, Jun. 2025.

[28] D. Yeke, Y. Zhou, L. Y. Lin, H. Cai, A. Bianchi, and Z. B. Celik, "RoboJailBench: Benchmarking Adversarial Attacks and Defenses in Embodied Robotic Agents," *arXiv preprint arXiv:2605.19328*, 2026.

[29] D. Yeke, E. S. Temirel, A. Shreekumar, B. Lee, D. Xu, and Z. B. Celik, "The Yes-Man Syndrome: Benchmarking Abstention in Embodied Robotic Agents," *arXiv preprint arXiv:2605.20544*, 2026.

[30] A. Chu et al., "NetSSM: Multi-Flow and State-Aware Network Trace Generation Using State-Space Models," *Proc. ACM Netw.*, vol. 4, no. CoNEXT1, pp. 1–24, 2026.

[31] S. Sonya, M. Zou, S. Vasan, C. Kruegel, G. Vigna, and D. Xu, "One Size Doesn't Fit All: A Dynamic Heterogeneous Learning Ensemble for Malware Family Classification," in *Proc. Eur. Symp. Res. Comput. Secur. (ESORICS)*, Cham, Switzerland: Springer Nature, 2025, pp. 233–253.

[32] Z. Xu et al., "Magpie: Alignment Data Synthesis from Scratch by Prompting Aligned LLMs with Nothing," *Magpie*. https://magpie-align.github.io/ (accessed Jun. 1, 2026).

[33] Z. Xu et al., "Magpie: Alignment Data Synthesis from Scratch by Prompting Aligned LLMs with Nothing," *arXiv preprint arXiv:2406.08464*, 2024.

[34] Z. Xu, A. M. Soria, S. Tan, A. Roy, A. S. Agrawal, R. Poovendran, and R. Panda, "TOUCAN: Synthesizing 1.5M Tool-Agentic Data from Real-World MCP Environments," presented at the *ICLR 2026 Workshop on Scaling Post-training for LLMs (SPOT)*, Rio de Janeiro, Brazil, Apr. 2026. [Online]. Available: https://arxiv.org/abs/2510.01179

[35] S. Ullah, P. Balasubramanian, W. Guo, A. Burnett, H. Pearce, C. Kruegel, G. Vigna, and G. Stringhini, "From CVE Entries to Verifiable Exploits: An Automated Multi-Agent Framework for Reproducing CVEs," *arXiv preprint arXiv:2509.01835*, 2025.

[36] F. Uccello, I. Karim, K. S. Mubasshir, E. Bertino, and S. Nadjm-Tehrani, "Mission Explainable: From Feature Attribution to Mitigation in 5G Anomaly Detection," in *Proc. IEEE Int. Conf. Comput., Netw. Commun. (ICNC)*, Maui, HI, USA, Feb. 2026.

[37] E. Debenedetti et al., "AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents," *arXiv preprint arXiv:2406.13352*, 2024. [Online]. Available: https://arxiv.org/abs/2406.13352

[38] Amazon, "Pushing the Boundaries of Secure AI: Winners of the Amazon Nova AI Challenge," *Amazon Science*, 2025. https://www.amazon.science/nova-ai-challenge/pushing-the-boundaries-of-secure-ai-winners-of-the-amazon-nova-ai-challenge (accessed Jun. 1, 2026).

[39] AIxCC, "DEF CON 33," *AI Cyber Challenge*. https://aicyberchallenge.com/def-con-33/ (accessed Jun. 1, 2026).

[40] F. Jiang, F. Ma, Z. Xu, Y. Li, B. Ramasubramanian, L. Niu, B. Li, X. Chen, Z. Xiang, and R. Poovendran, "SoSBench: Benchmarking Safety Alignment on Scientific Knowledge," presented at the *NeurIPS 2025 Workshop on Socially Responsible and Trustworthy Foundation Models*, San Diego, CA, USA, Dec. 2025.

[41] Q. Wu, Z. Chen, W. Corcoran, M. Sra, and A. Singh, "GraphEval36K: Benchmarking Coding and Reasoning Capabilities of Large Language Models on Graph Datasets," in *Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics (NAACL)*, Albuquerque, NM, USA, Apr.–May 2025.

[42] Y. Feng et al., "Visual Aesthetic Benchmark: Can Frontier Models Judge Beauty?," *arXiv preprint arXiv:2605.12684*, 2026.

[43] A. Bhardwaj, N. Diwan, and G. Wang, "Beyond BeautifulSoup: Benchmarking LLM-Powered Web Scraping for Everyday Users," *arXiv preprint arXiv:2601.06301*, 2026.

[44] DIMACS, "REU Program at DIMACS," *Rutgers University*. https://reu.dimacs.rutgers.edu/ (accessed Jun. 1, 2026).

[45] Carnegie Mellon University, "picoCTF - Free Computer Science Education Program," *picoCTF*. https://picoctf.org/ (accessed Jun. 1, 2026).

[46] SANS Institute, "SANS Skills Quest," *SANS Institute*. https://www.sans.org/cyber-ranges/skills-quest (accessed Jun. 1, 2026).

[47] Cyber Sweden Conference, "Cyber Sweden," *Cyber Sweden Conference*. https://cyberswedenconf.se/ (accessed Jun. 1, 2026).

[48] IBM Research, "Toucan: A New Goldmine for Tool-Calling AI Agents," *IBM Research Blog*. https://research.ibm.com/blog/toucan-for-tool-calling (accessed Jun. 1, 2026).

[49] ReversingLabs, "CVE Genie and the Future of Vulnerability Management," *ReversingLabs Blog*. https://www.reversinglabs.com/blog/cve-genie-vulnerability-management (accessed Jun. 1, 2026).

[50] Y. Li et al., "JobBench: Aligning Agent Work With Human Will," *arXiv preprint arXiv:2605.26329*, 2026.

[51] M. Wong, "Chatbots Are Becoming Really, Really Good Criminals," *The Atlantic*, Nov. 2025. https://www.theatlantic.com/technology/2025/11/anthropic-hack-ai-cybersecurity/685061/ (accessed Jun. 1, 2026).

[52] Sunblaze UCB, "cybergym," *GitHub*. https://github.com/sunblaze-ucb/cybergym (accessed Jun. 1, 2026).

[53] SCSP, "Special Competitive Studies Project," *SCSP*. https://www.scsp.ai/ (accessed Jun. 1, 2026).

[54] CoSAI, "Coalition for Secure AI," *CoSAI*. https://coalitionforsecureai.org/ (accessed Jun. 1, 2026).

[55] Shellphish, "Agentish," *Agentish*. https://agentish.io/ (accessed Jun. 1, 2026).

[56] ACM, "ACM Elects New Officers for 2026-2028 Term," *ACM Media Center*, May 20, 2026. https://www.acm.org/media-center/2026/may/2026-election-results (accessed Jun. 1, 2026).

[57] Perplexity AI, "Announcing the Secure Intelligence Institute," *Perplexity Blog*, Mar. 9, 2026. https://www.perplexity.ai/hub/blog/announcing-the-secure-intelligence-institute (accessed Jun. 1, 2026).

[58] Z. Chen, M. Kang, and B. Li, "ShieldAgent: Shielding Agents via Verifiable Safety Policy Reasoning," in *Proc. 42nd Int. Conf. Mach. Learn. (ICML)*, PMLR vol. 267, 2025.

Appendix A: Glossary and Acronyms

ACTION – AI Institute for Agent-based Cyber Threat Intelligence and OperatioN

Agent – We define an intelligent security agent as an [AI-powered] entity that performs one or more cybersecurity functions, with reasoning, learning, and collaboration (human-agent and inter-agent) capabilities.

AI – Artificial intelligence

AIVO – AI Institutes Virtual Organization

AIxCC – DARPA's AI Cyber Challenge

AP – Advanced Placement

APT – Advanced Persistent Threat

ARPA-H – Advanced Research Projects Agency for Health

ASU – Arizona State University

AVATAR – Autonomous Vulnerability Analysis, Triaging, And Repair

AWS – Amazon Web Services

C4DT – Center for Digital Trust, a research unit at the Swiss Federal Institute of Technology Lausanne (EPFL)

CAE – Centers for Academic Excellence in Cybersecurity (NSA program)

CDET – Cyber Defense Education and Training (CISA program)

CISA – Cybersecurity and Infrastructure Security Agency

CISO – Chief Information Security Officer

CPS – Cyber-physical Systems

CPSR – Cyber-Physical Systems Resilience (NITRD working group)

CRS – Cyber Reasoning System

CSET – Center for Security and Emerging Technology, research group at Georgetown University

CSIA – Cyber Security and Information Assurance (NITRD working group)

CTF – Capture the Flag (security exercise)

CTE – Career Technical Education

CTI – Cyber Threat Intelligence

CVE – Common Vulnerabilities and Exposures

CVE-GENIE – Automated multi-agent system for reproducing verifiable exploits from CVE entries

CVEX – CVE Exploitation framework (ACTION tool for vulnerability exploitation research)

DARPA – Defense Advanced Research Projects Agency

DHS – Department of Homeland Security

DHS S&T – Department of Homeland Security Science and Technology

DIMACS – Center for Discrete Mathematics and Theoretical Computer Science (Rutgers University)

DNNs – Deep Neural Networks

DOD – Department of Defense

EAB – ACTION Institute External Advisory Board (see management structure)

EC – ACTION Institute Executive Committee (part of management structure)

ECE – Electrical and Computer Engineering

EDRN – Early Detection Research Network (National Cancer Institute)

EPFL – Swiss Federal Institute of Technology Lausanne

GAO – Government Accountability Office

GATE – Globally Accessible Testing Environment

GRIT – Ground-breaking Research/Innovative Technology, an outreach program at UCSB where scientists present research to the community

HBCU – Historically Black College or University

iCTF – International Capture the Flag

IFG – Information flow graph

JSM – Joint Statistical Meeting (Institute for Mathematical Statistics)

KG – Knowledge graph

LLM – Large Language Model

LP – Linear Programming

MABEL – Malware Analysis Benchmark for AI & ML, a dataset released by ACTION researchers

MCA – Multi-Campus Agreement, a subaward agreement where the prime and sub are part of the same university, but on different campuses

ML – Machine learning

MOOC – Massive Open Online Course

MSB – Swedish Civil Contingencies Agency

MSI – Minority Serving Institution

NAIRR – National Artificial Intelligence Research Resource Pilot

NCI – National Cancer Institute

NCSA – National Center for Supercomputing Applications

NEH – National Endowment for the Humanities

NICCS – National Initiative for Cybersecurity Careers and Studies

NICT – National Institute of Information and Communications Technology (Japan)

NIST – National Institute of Standards and Technology

NITRD – Networking and Information Technology Research and Development

NLP – Natural Language Processing

NSA – National Security Agency

PCS – Predictability-computability-stability framework

PLC – Programmable Logic Controller

PPL – Probabilistic Programming Language

RCR – Responsible Conduct of Research, a training through CITIprograms on proper research methods that are both reproducible and ethical

RDI – Berkeley Center for Responsible Decentralized Intelligence

REU – Research Experience for Undergraduates

RFP – Request for Proposals

RL – Reinforcement Learning

R&R – Response and Recovery

ROXAS – Reasoning Over eXplained AnomalieS

SACNAS – Society for the Advancement of Chicanos/Hispanics and Native Americans in Science

SAC – ACTION Institute Student Advisory Council (see management structure)

SAIL – Summit for AI Institute Leadership

SANS – SysAdmin, Audit, Network, and Security Institute

SAAS – Software as a Service

SAT – Satisfiability problems, which are core challenges in software engineering and have applications in tasks like automated testing and program verification

SBIR – Small Business Innovation Research

SCSP – Special Competitive Studies Project

SDP – Semidefinite Programming

SES – Socioeconomic Status

S&IP / SIP – Strategic and Implementation Plan

SLED – State, Local, and Education (organizations)

SOC – Security Operations Center

SoK – Systematization of Knowledge

SPHERE – Scalable, Programmable, High-fidelity testbed for cybersecurity research (USC/Utah)

SST – School for Scientific Thought, an educational program for Broadening Participation, run by UCSB's Center for Science and Engineering Partnerships (CSEP) and Early Academic Outreach Program (EAOP)

SWOT – Strengths, Weaknesses, Opportunities, and Threats

TI – Threat Intelligence

TOUCAN – Tool-use dataset synthesizing 1.5 million data points from real-world Model Context Protocols, developed in collaboration with IBM

TTP – Tactics, Techniques, and Procedures

TTP-T – NSF Technology Transfer and Translation program

UC – University of California

UCSB – University of California, Santa Barbara

UIUC – University of Illinois, Urbana-Champaign

VAB – Visual Aesthetic Benchmark

VDS – Veridical Data Science

VFL – Vertical Federated Learning

VFLAIR – Vertical Federated Learning framework used for evaluating adversarial evasion attacks

VPC – Virtual Private Cloud

WEBRR – Web-based Response and Recovery tool (ACTION research output)

WiCS – Women in Computer Science

WiCyS – Women in Cybersecurity

WWU – Western Washington University

XAI – Explainable AI

Appendix B: Highlights of Significant Achievement and Impact

Highlight 1: Lifelong Detection and Learning of New Threats

Bing Liu (University of Illinois Chicago), Xiaoming Huo and Wenke Lee (Georgia Tech)

Outcome/Accomplishment

Today's cybersecurity AI systems are trained once and then frozen — unable to adapt as attackers invent new methods. Researchers from the University of Illinois Chicago and Georgia Tech, combining expertise in machine learning theory, statistics, and systems security, are developing AI that learns continuously and autonomously. These systems discover new threats on their own, incorporate that knowledge, and correct mistakes — all without requiring human engineers to intervene or retrain the system from scratch. This required a depth and breadth of expertise that no single investigator or institution could have provided alone, and the tight integration of theoretical machine learning with operational security knowledge is what made it possible.

Impact and Benefits

Cyberattacks cost the global economy trillions of dollars annually, and attackers consistently hold one decisive advantage: they change tactics faster than defenders can update their systems. Every new attack technique triggers an exhausting cycle of data collection, labeling, and retraining that can take weeks or months while threats continue to cause damage. This research directly attacks that asymmetry. By enabling AI to detect when something new is happening and adapt accordingly — and by developing "unlearning" algorithms that allow deployed systems to surgically correct errors without a full, costly retraining — the team has laid the groundwork for security systems that grow more capable over time rather than becoming stale. The scientific advances made here extend beyond cybersecurity to any domain where AI must operate reliably in a world that keeps changing after deployment.

Key Achievement

During the reporting year, the team advanced autonomous lifelong learning algorithms that detect and incorporate new threats without human-labeled data, and developed validated unlearning methods that correct erroneous knowledge in deployed systems

without full retraining. A paper demonstrating these results is currently under peer review, with two additional manuscripts to be submitted imminently.

Primary ACTION Research Thrust

Thrust AI-1: Learning and Reasoning with Domain Knowledge

References: [13-15]

Highlight 2: Adaptive Design of AI Assistance to Enhance Human Decision Making

Ming Yin (Purdue University) and Ambuj Singh (UC Santa Barbara)

Outcome/Accomplishment:

Researchers from Purdue and UC Santa Barbara combined behavioral science, human-computer interaction, and AI systems expertise to develop AI assistance that adapts to the individual, to the context, and to the specific risks of the moment.

Impact and Benefits

One of the most dangerous problems in modern cybersecurity is the human analyst who works alongside AI systems and routinely develops over-reliance, dismissing threats the AI doesn't flag, and trusting recommendations (even when they are wrong).

This research produced two validated frameworks addressing this problem. The first introduces an AI system that intelligently switches between building appropriate analyst trust and aggressively catching threats human intuition would miss. The second shows that AI designed to challenge analyst reasoning, rather than simply advise, measurably reduces dangerous over-reliance.

As AI takes on larger roles in national security operations, keeping human analysts genuinely engaged and critically thinking is paramount. This research provides a concrete, validated path to achieving such a goal.

Key Achievement

Research results demonstrate that AI-driven critical thinking challenges reduce analyst over-reliance while preserving team performance. During the reporting year, the team

published two peer-reviewed papers: one at AAAI 2026 introducing the theoretically grounded adaptive AI ensemble with Rational Routing Shortcut mechanism, and one at CHI 2026. The latter also received an honorable mention award.

Primary ACTION Research Thrust

Thrust AI-2: Human-Agent Teaming

References: [2, 3]

Highlight 3: Learning with Regret Minimization for Cybersecurity Applications

João Hespanha (UC Santa Barbara) and Radha Poovendran (University of Washington)

Outcome/Accomplishment

Cyber adversaries are not static. They watch, adapt, and probe for weaknesses in whatever defense they face. Researchers from UC Santa Barbara and the University of Washington combined expertise in control theory, online learning, reinforcement learning, and cybersecurity to develop a new class of algorithms that provide mathematical guarantees on defensive performance regardless of how sophisticated or unpredictable the adversary turns out to be. A significant extension of the work into protecting critical infrastructure such as power grids, water systems, and autonomous transportation emerged directly from this interdisciplinary partnership.

Impact and Benefits

How do you make good decisions when you don't know how smart your opponent is? In network security this question is constant, and being wrong about the answer is costly. Poovendran and Hespanha's research developed algorithms that automatically sense how adversarially variable the current threat environment is and adjust strategy accordingly, thereby achieving provably optimal performance across the spectrum of predictable, slowly-evolving threats to rapidly-shifting, fully-adaptive adversaries. A new structural measure, called polyhedral instability, gives defenders a quantifiable way to characterize this variability and thus advances the science of adversarial machine learning. The extension to cyber-physical systems carries major national significance. As power plants, water facilities, and transportation networks become

more AI-dependent, their exposure to stealthy cyberattacks grows. This work provides both the mathematical framework for understanding those vulnerabilities and the algorithmic tools to defend against them, even when attackers have inside knowledge of the systems they are targeting.

Key Achievements

During the reporting year, the team

- Published region-aware learning algorithms with provably optimal regret bounds spanning the full range of adversarial variability
- Demonstrated and formally proved how standard reinforcement learning policies can be exploited in large state-space security problems and introduced an opponent- informed defense
- Extended the framework to cyber-physical systems, establishing Nash equilibrium security policies against worst-case stealthy attackers.

These results appeared across multiple peer-reviewed venues.

Primary ACTION Research Thrust

AI-4: Strategic Gaming and Tactical Planning

References: [24-27]

Highlight 4: Leto: Reinforcement Learning to Optimize Intrusion Response Against Network Threats

Gang Wang (University of Illinois Urbana-Champaign) and Christopher Kruegel (UC Santa Barbara)

Outcome/ Accomplishment

Researchers from the University of Illinois Urbana-Champaign and UC Santa Barbara combined expertise in network security, reinforcement learning, and large-scale systems to build an AI that not only detects intrusions, but also automatically determines and executes the optimal response, blocking attackers while keeping legitimate traffic flowing without interruption. The result is a system that learns from its own experience in a real network environment, improving its response policy over time without requiring human engineers to manually tune its rules.

Impact and Benefits

Hospitals, universities, financial institutions, and government agencies face a relentless stream of malicious network traffic. The standard response is a combination of human analysts and rule-based blocking systems, both of which are slow, expensive, and imperfect. Human analysts experience fatigue and cannot scale. Rule-based systems cannot adapt to novel attack patterns. The cost of getting this wrong runs into billions of dollars annually across the economy. *Leto* addresses this problem directly by automating the response decision itself, not just the detection. By learning which attacker behaviors are truly dangerous and which alerts are false alarms, the system eliminates the manual tuning burden that makes current intrusion response systems so costly to operate.

Critically, *Leto* was designed from the outset to handle evasive attackers, making it effective precisely in the scenarios where conventional systems fail. The national security implications are significant: as adversaries become more sophisticated, the ability to respond automatically, accurately, and at machine speed becomes a strategic necessity.

Key Achievements

During the reporting year, the team evaluated *Leto* in a real-world pilot deployment and tested whether its learned response policy holds up against threats that emerged after training. A manuscript describing the results is under review.

Primary ACTION Research Thrust

SEC-2: Intelligent Agents for Detection

Highlight 5: Shield Agent: Unified and Knowledge-Enabled Guardrail Agents to Protect Diverse Agentic Systems

David Wagner & Dawn Song (UC Berkeley) and Bo Li (University of Illinois Urbana-Champaign)

Outcome/ Accomplishment

ACTION researchers from UC Berkeley and the University of Illinois Urbana-Champaign developed *Shield Agent*, a unified guardrail framework that protects AI agents against prompt injection.

Impact and Benefits

The stakes of getting AI agent security wrong are extraordinarily high and growing rapidly. An AI agent with access to a user's email, calendar, financial accounts, and computer can cause devastating harm if hijacked. The attacks that hijack them are often invisible, requiring no hacking of the underlying model. It only takes a cleverly worded instruction embedded in content the agent was going to read anyway. This attack, known as prompt injection, has been identified by industry practitioners as the single greatest security threat facing AI agent systems today.

As organizations race to deploy AI agents to automate customer service, software development, medical records management, and national security workflows, the absence of robust guardrails creates systemic risk at a scale the cybersecurity industry has not previously encountered.

Shield Agent builds a protection layer that travels with the agent rather than depending on the security of every piece of content the agent might encounter. The dual-agent privilege-separation framework, in which one agent handles untrusted inputs while a separate, authority-holding agent controls all security-sensitive actions, mirrors the principle of least privilege that underlies decades of sound security engineering. The breadth of applicability is perhaps the most significant aspect of this work: AI agents across many industries can benefit from the architectural patterns Shield Agent establishes.

Key Achievements:

Shield Agent's prompt injection defense demonstrates that separating trust boundaries between agents can substantially improve security while maintaining full task utility.

Primary ACTION Research Thrust

SEC-2: Intelligent Agents for Detection

Builds on the published ShieldAgent paper [58].

Highlight 6: Klleem: Multi-Level, Targeted Vulnerability Analysis in Large Networks

Wenke Lee (Georgia Tech) and Christopher Kruegel (UC Santa Barbara)

Outcome / Accomplishment

Researchers from Georgia Tech and UC Santa Barbara combined expertise in large language models, game-theoretic threat modeling, program analysis, and fuzzing to build *Klleem*, a Dynamic Profiling Agent that works the way a skilled human security researcher would, but at machine speed and scale. Rather than scanning everything equally, *Klleem* first uses an LLM-guided Monte Carlo simulation to reason about which attack paths across a network are most likely to be exploited by a strategic adversary, focusing attention where it matters most. *Klleem* then directs deep, targeted code analysis techniques (fuzzing and symbolic execution) precisely at those high-risk regions, uncovering vulnerabilities that broad-coverage tools routinely miss.

Impact and Benefits

Software vulnerabilities are the entry point for the vast majority of serious cyberattacks, from ransomware that shuts down hospitals, to espionage campaigns that steal national secrets. The challenge is not that we lack tools to find vulnerabilities. Rather, it is that the tools we have generate too much noise to be actionable in large, real-world systems. They are blind to the strategic question of which vulnerabilities an attacker would actually exploit. *Klleem* contributes two important changes. First, by prioritizing analysis based on what a strategic adversary would target, rather than what is merely statistically anomalous, it aligns the defender's effort with the attacker's incentives. Second, by integrating LLM-based semantic understanding of code structure into the fuzzing process, it generates far higher-quality test inputs than mutation-based fuzzers that operate without understanding what the code is trying to do.

The practical result is faster discovery of more serious vulnerabilities with less wasted analyst time. This is a combination that directly improves the security posture of any organization that deploys it. At a national level, the ability to systematically identify the highest-risk vulnerabilities in critical software and infrastructure before adversaries exploit them represents a meaningful shift in the asymmetry between attackers and defenders.

Key Achievements

During the reporting year, the team broadened *Klleem's* evaluation to additional real-world programs and developed new techniques for guiding code analysis toward high-risk regions. A manuscript describing the results is under review.

Primary ACTION Research Thrust

Thrust SEC-1: Intelligent Agents for Threat/Vulnerability Assessment
Thrust AI-4: Strategic Gaming and Tactical Planning

Highlight 7: iCTF 2026 “Battle of the Agents” and the Agentish Platform

Giovanni Vigna (UC Santa Barbara), Saastha Vasan (PhD Student, UC Santa Barbara), and many other PhD students from across ACTION

Outcome / Accomplishment

Under the theme "Battle of the Agents," iCTF 2026 became the first major international CTF competition to incorporate agentic challenges: security tasks that directly reflect the growing role of AI and autonomous systems in real-world cybersecurity. Running five days across separate high school and undergraduate divisions, the competition drew participants from the United States, India, Germany, and Mexico. Research breakthroughs across ACTION's security and AI thrusts directly informed the challenge designs and inspired the launch of Agentish, a new visual platform developed chiefly by Saastha Vasan, a PhD student at UCSB. Agentish allows students and researchers to design, test, and deploy autonomous AI agent pipelines through an intuitive node-based graph editor, with MCP tool integration and real CTF challenge support.

Impact and Benefits

The cybersecurity workforce shortage is a national security concern, made more acute each year as AI becomes central to both attack and defense. Most training programs have not yet caught up to the reality that tomorrow's security professionals must understand how to build, attack, and defend AI agent systems, and not just traditional networks. iCTF 2026 puts agentic security challenges directly in front of the next generation of practitioners at the moment career trajectories are being set. The Agentish platform amplifies this impact by giving students a reusable, accessible environment for building autonomous agent pipelines. These are skills directly transferable to the most in-demand roles in the security industry. Its open design means it can be adopted by educators and researchers well beyond the ACTION community, creating durable workforce development infrastructure rather than a one-time event.

Key Achievement

iCTF 2026 ran March 9–13, 2026, featuring 64 total challenges including 16 agentic challenges. This was a first for any major international CTF.

The high school division generated 310 flag submissions. Prizes went to Thomas Jefferson High School for Science and Technology (1st, $3,000), the US Cyber Games Pipeline Program (2nd, $2,000), and Eastern Technical High School (3rd, $1,000).

The undergraduate division recorded 284 correct submissions across 2,318 attempts, with top finishers from institutions in India, Germany, and Mexico.

Simultaneously, the Agentish platform launched publicly at agentish.io and is now available to the global security research and education community.

Primary ACTION Research Thrust

Education and Outreach

Reference: [55]

Highlight 8: CVE-Genie: Automated End-to-End Vulnerability Reproduction

Saad Ullah (Boston University), with ACTION researchers from UC Santa Barbara and also Arizona State University and University of New South Wales

Outcome / Accomplishment

CVE-Genie is a multi-agent AI system that takes a raw CVE record, automatically gathers the relevant source code and vulnerability data, rebuilds the exact vulnerable software environment, generates working exploit code, and then independently verifies that the exploit actually works, all without human intervention. A collaborative pipeline of specialized agents achieves reliable results across an enormous variety of real-world vulnerability types.

Impact and Benefits

The window between when a vulnerability is publicly disclosed and when attackers weaponize it has shrunk sharply in recent years. Security teams operating on manual patch cycles and ad hoc processes simply cannot keep pace. CVE-Genie directly addresses this asymmetry by automating the production of verifiable, reproducible exploit code at scale and at a cost of just $2.77 per vulnerability. The low cost makes it possible to generate the high-quality security benchmarks that the entire field depends on for evaluating tools, validating patches, and training AI-based detection systems. In a comprehensive evaluation on 841 real-world vulnerabilities published in 2024 and 2025, CVE-Genie successfully reproduced working exploits for 428, a 51% success rate across 22 programming languages and 267 projects. Industry experts described the result as a major leap forward, predicting it will transform vulnerability management from theoretical to empirical, based on provable exploitability. The tool has also prompted serious dual-use discussion. By lowering the barrier to exploit generation, CVE-Genie raises urgency for defenders to adopt it before attackers do.

Key Achievements

CVE-Genie was published on arXiv, publicly released as open-source software at github.com/BUseclab/cve-genie, and covered by ReversingLabs, a leading industry security intelligence firm, in a widely circulated blog post. The system's EAGER design framework — Exploit generation, Assessment, Generalization, End-to-end automation, and environment Rebuilding — and its four-component pipeline (processor, builder, exploiter, and CTF verifier) represent a concrete, deployable advance in automated vulnerability science that directly supports ACTION's research thrusts.

Primary ACTION Research Thrust

SEC-1: Intelligent Agents for Risk/Vulnerability Assessment, SEC-3: AI-driven Cyber Threat Intelligence, AI-3: Multi-Agent Collaboration and AI-1: Learning and Reasoning

References: [35, 49]

Highlight 9: ARTIPHISHELL: Autonomous Cyber Reasoning at Machine Speed

Giovanni Vigna (UC Santa Barbara), with researchers from Arizona State University and Purdue University

Outcome / Accomplishment

At DEF CON 33 in Las Vegas, Team Shellphish, which shares membership with ACTION, debuted ARTIPHISHELL, an autonomous Cyber Reasoning System built directly on foundational research from the ACTION AI Institute. As one of only seven finalist teams globally in the DARPA AI Cyber Challenge (AIxCC), Shellphish demonstrated that ACTION's core research architecture of intelligent, cooperating security agents can be translated into a deployed system that finds and fixes real software vulnerabilities at machine speed. ARTIPHISHELL moved well beyond simple LLM prompting by deploying an ensemble of over 60 specialized agents, allowing the system to perform complex triage and generate verified patches.

Impact and Benefits

The world's critical software infrastructure underpinning banking systems, hospital networks, power grids, and public communications contains an unknown but enormous number of unpatched vulnerabilities. Human security teams cannot find and fix them fast enough. Manual vulnerability response typically costs thousands of dollars per flaw, and takes weeks. The seven AIxCC finalist systems, including ARTIPHISHELL, collectively identified 77% of synthetic vulnerabilities and discovered 18 previously unknown zero-day flaws in the challenge codebases. These are real vulnerabilities that no human had found before. Across the finalist systems, DARPA reported an average patch time of just 45 minutes and an average cost of approximately $152 per vulnerability — a reduction of one to two orders of magnitude compared to current manual practice. These results suggest that, for the first time, autonomous AI systems have demonstrated the ability to find and fix previously unknown real-world software flaws at a cost and speed that could plausibly keep pace with the scale of the problem.

Key Achievements

ARTIPHISHELL is a direct application of ACTION's use-inspired research framework, showing that the institute's research can produce technology that reaches practitioners and helps protect national infrastructure.

Primary ACTION Research Thrust

SEC-1: Intelligent Agents for Risk/Vulnerability Assessment
SEC-4: Intelligent Agents for Response and Recovery
AI-1: Learning and Reasoning

AI-3: Multi-Agent Collaboration

Reference: [39]

Highlight 10: PurpCorn-PLAN: Winning the Amazon Nova AI Challenge

Gang Wang and team, including Hadjer Benkraouda (UIUC)

Outcome / Accomplishment

In the first-ever Amazon Nova AI Challenge, a head-to-head adversarial tournament unlike any prior AI safety evaluation, Team PurpCorn-PLAN from the University of Illinois Urbana-Champaign, including ACTION researcher Hadjer Benkraouda and ACTION co-PI Gang Wang, won the defending team competition by outperforming international competitors. PurpCorn-PLAN's success drew directly on ACTION's research agenda; the team's approach to reasoning-based safety alignment, adversarial robustness, and human-agent interaction translated directly into a winning competitive system.

Impact and Benefits

The security of AI coding assistants is not an abstract concern. These tools are already being deployed to write code in healthcare systems, financial platforms, transportation infrastructure, and national security applications. When a coding AI can be tricked into generating malicious or vulnerable code, the consequences flow directly into production software used by millions of people. The Amazon Nova AI Challenge surfaced several findings with immediate industry relevance: multi-turn attack planning proved far more effective than single-turn jailbreaks, demonstrating that safety systems must account for progressive adversarial escalation across a conversation, not just isolated prompts. Reasoning-based safety alignment, as used by PurpCorn-PLAN, successfully prevented vulnerabilities without degrading code generation utility, resolving what had previously been assumed to be an unavoidable tradeoff between safety and capability.

Synthetic data generation emerged as a scalable path to continuously improving model safety without prohibitive human red-teaming costs. These findings, validated under

genuine adversarial pressure against an 8B model built by Amazon specifically for the competition, go well beyond what typical academic research environments can test.

Key Achievement

Team PurpCorn-PLAN demonstrated reasoning-based guardrails, policy optimization, and vulnerability fixers applied to Amazon's custom 8B coding model, preventing adversarial jailbreaks while maintaining full code generation utility. The team presented their findings at the Amazon Nova AI Summit in Seattle.

Primary ACTION Research Thrusts

SEC-2: Intelligent Agents for Detection
SEC-4: Intelligent Agents for Response and Recovery
AI-3: Multi-Agent Collaboration
AI-4: Strategic Gaming and Tactical Planning

Reference: [38]

Highlight 11: TOUCAN: Supercharging Tool-Calling AI Agents

Radha Poovendran (University of Washington), in collaboration with IBM Research

Outcome / Accomplishment

Building on their earlier work on synthesizing alignment data (Magpie), ACTION researchers led by Radha Poovendran at the University of Washington, in collaboration with IBM Research, created TOUCAN, the largest publicly available dataset of tool-calling scenarios to date. TOUCAN contains 1.5 million real-world task trajectories synthesized from nearly 500 authentic Model Context Protocol (MCP) servers, the standardized interfaces through which AI agents connect to the world's APIs. TOUCAN covers more than 2,000 tools and includes multi-turn and parallel tool-calling scenarios.

Impact and Benefits

The ability of AI agents to reliably use software tools is the capability that transforms a language model from a chatbot into a genuine autonomous assistant. Every major industry is now racing to deploy AI agents for exactly these kinds of tasks, but the

quality of those agents is directly limited by the quality of the training data available. TOUCAN breaks this bottleneck open. Small open-source models fine-tuned on TOUCAN data achieved performance gains of up to seven percentage points on leading agentic benchmarks, and a TOUCAN-tuned Qwen-2.5-32B model improved by nearly nine percentage points on the Berkeley Function Calling Leaderboard. The dataset became a top trending resource on Hugging Face almost immediately upon release, attracting immediate community attention. By releasing TOUCAN as open-source, Poovendran's team and IBM have provided a public good that will accelerate the development of capable, open-source AI agents across healthcare, finance, cybersecurity, and any other domain where reliable tool-using agents are needed.

Key Achievement

During the reporting year, TOUCAN was released publicly on Hugging Face (1.5 million trajectories), its pre-print paper was posted on arXiv (arxiv.org/abs/2510.01179), and the dataset was open-sourced on GitHub. IBM Research published a dedicated blog post by science writer Kim Martineau hailing TOUCAN as a "new goldmine for tool-calling AI agents," and the dataset immediately became a top trending resource on Hugging Face upon release. In benchmark evaluations, TOUCAN-tuned open-source models outperformed larger closed-source models on BFCL V3 and achieved state-of-the-art results for their size on MCP-Universe, two leading benchmarks for agentic tool use, demonstrating that the quality and realism of TOUCAN's training data translates directly into measurable, state-of-the-art agent capability gains.

Primary ACTION Research Thrust

Thrust AI-1: Learning and Reasoning with Domain Knowledge
Thrust AI-3: Multi-Agent Collaboration
Thrust SEC-4: Intelligent Agents for Response and Recovery

References: [32-34, 48]

Appendix C: List of Institute Projects

Project: Modeling and Explaining Dynamic Graphs Thrust AI-1: Learning and Reasoning with Domain Knowledge Research/Technology -- Direct Support			
Position Title	Name	Department	Institution
Activity Co-Leader	Ambuj Singh	Computer Science	UC Santa Barbara
Activity Co-Leader	Jie Gao	Computer Science	Rutgers University
Postdoctoral Research	Cheng Xin	Computer Science	Rutgers University
Graduate Student	Callie Sardina	Computer Science	UC Santa Barbara
Graduate Student	Eren Polat	Computer Science	UC Santa Barbara
Graduate Student	Daniel Baumgartner	Computer Science	Rutgers University

Project: LLMs for Cybersecurity Time Series Reasoning Thrust AI-1: Learning and Reasoning with Domain Knowledge Research/Technology -- Direct Support			
Position Title	Name	Department	Institution
Activity Co-Leader	David Evans	Computer Science	University of Virginia
Activity Co-Leader	Sewoong Oh	Computer Science and Engineering	University of Washington
Graduate Student	Daniel Baumgartner	Computer Science	Rutgers University

Project: Lifelong Detection and Learning of New Threats on the Fly: Theory and Algorithms Thrust AI-1: Learning and Reasoning with Domain Knowledge Research/Technology -- Direct Support			
Position Title	Name	Department	Institution
Activity Co-Leader	Bing Liu	Computer Science	UI Chicago
Activity Co-Leader	Xiaoming Huo	Statistics	Georgia Tech
Faculty Researcher	Wenke Lee	Computer Science	Georgia Tech
Graduate Student	Saleh Momeni	Computer Science	UI Chicago

Project: Adaptive Design of AI Assistance to Enhance Human Decision Making Thrust AI-2: Human-Agent Interaction Research/Technology -- Direct Support			
Position Title	Name	Department	Institution
Activity Co-Leader	Ming Yin	Computer Science	Purdue University
Activity Co-Leader	Ambuj Singh	Computer Science	UC Santa Barbara
Graduate Student	Hasan Amin	Computer Science	Purdue University

Project: Design LLM-based Agents to Enhance Human-Agent Collective Decision-Making Thrust AI-2: Human-Agent Interaction Research/Technology -- Direct Support			
Position Title	Name	Department	Institution
Activity Co-Leader	Ming Yin	Computer Science	Purdue University
Activity Co-Leader	Ambuj Singh	Computer Science	UC Santa Barbara
Graduate Student	Abed K. Musaffar	Computer Science	UC Santa Barbara
Graduate Student	Rasta Tadayontahmasebi	Computer Science	UC Santa Barbara

Project: Adversary-aware Agent Interaction, Reasoning, and Learning Thrust AI-3: Multi-Agent Collaboration Research/Technology -- Direct Support			
Position Title	Name	Department	Institution
Activity Co-Leader	Jie Gao	Computer Science	Rutgers University
Activity Co-Leader	Ninghui Li	Computer Science	Purdue University
Graduate Student	Kevin Lu	Computer Science	Rutgers University
Graduate Student	Kaiyuan Zhang	Computer Science	Purdue University
Faculty Researcher	Elisa Bertino	Computer Science	Purdue University
Postdoctoral Scholar	Cheng Xin	Computer Science	Rutgers University

Project: MedAgent: A Secure Multi-Agent System for Healthcare Thrust AI-3: Multi-Agent Collaboration Research/Technology -- Direct Support			
Position Title	Name	Department	Institution
Activity Co-Leader	Bo Li	Computer Science	UI Urbana-Champaign
Activity Co-Leader	Bin Yu	Statistics	UC Berkeley
Graduate Student	Aaron J. Li	Statistics	UC Berkeley

Project: Learning with regret minimization for cybersecurity applications Thrust AI-4: Strategic Gaming and Tactical Planning Research/Technology -- Direct Support			
Position Title	Name	Department	Institution
Activity Co-Leader	Radha Poovendran	Electrical and Computer Engineering	University of Washington
Activity Co-Leader	João Hespanha	Electrical and Computer Engineering	UC Santa Barbara
Graduate Student	Sean Anderson	Electrical and Computer Engineering	UC Santa Barbara
Graduate Student	Madeline Blischke	Electrical and Computer Engineering	UC Santa Barbara
Graduate Student	Fengqing Jiang	Electrical and Computer Engineering	University of Washington

Project: End-to-end Binary Reverse Engineering via LLM Reasoning and Tool Use Thrust SEC-1: Intelligent Agents for Threat/Vulnerability Assessment Research/Technology -- Direct Support			
Position Title	Name	Department	Institution
Activity Co-Leader	Gang Wang	Computer Science	University of Illinois at Urbana-Champaign
Activity Co-Leader	Berkay Celik	Computer Science	Purdue University
Graduate Student	Doguhan Yeke	Computer Science	Purdue University
Graduate Student	Pirouz Naghavi	Computer Science	University of Illinois at Urbana-Champaign
Faculty Researcher	Chris Kruegel	Computer Science	UC Santa Barbara

Project: Multi-Level, Targeted Vulnerability Analysis in Large Networks Thrust SEC-1: Intelligent Agents for Threat/Vulnerability Assessment Research/Technology -- Direct Support			
Position Title	Name	Department	Institution
Activity Co-Leader	Wenke Lee	Computer Science	Georgia Tech
Activity Co-Leader	Chris Kruegel	Computer Science	UC Santa Barbara
Graduate Student	Leyan Pan	Computer Science	Georgia Tech

Project: Context-Adaptive Malware Classification using Diversified Ensembles Thrust SEC-2: Intelligent Agents for Detection Research/Technology -- Direct Support			
Position Title	Name	Department	Institution
Activity Co-Leader	David Evans	Computer Science	University of Virginia
Activity Co-Leader	Chris Kruegel	Computer Science	UC Santa Barbara
Graduate Student	Minjun Elena Long	Computer Science	University of Virginia
Graduate Student	Stijn Pletinckx	Computer Science	UC Santa Barbara

Project: Shield Agent: Unified and Knowledge-Enabled Guardrail Agents to Protect Diverse Agentic Systems Thrust SEC-2: Intelligent Agents for Detection Research/Technology -- Direct Support			
Position Title	Name	Department	Institution
Activity Leader	David Wagner	Electrical Engineering and Computer Science	UC Berkeley
Faculty Researcher	Dawn Song	Electrical Engineering and Computer Science	UC Berkeley
Faculty Researcher	Bo Li	Computer Science	University of Illinois at Urbana-Champaign
Graduate Student	Zhaorun Chen	Computer Science	University of Chicago
Graduate Student	Mintong Kang	Computer Science	UI Urbana-Champaign

Project: Reinforcement Learning to Optimize Intrusion Response Against Network Threats (aka Leto) Thrust SEC-2: Intelligent Agents for Detection Research/Technology -- Direct Support			
Position Title	Name	Department	Institution
Activity Co-Leader	Gang Wang	Computer Science	UI Urbana-Champaign
Activity Co-Leader	Chris Kruegel	Computer Science	UC Santa Barbara
Graduate Student	Hadjer Benkraouda	Computer Science	UI Urbana-Champaign

Project: A Model-Based Study of Low-and-Slow APT Detection-Evasion Thrust SEC-2: Intelligent Agents for Detection Research/Technology -- Direct Support			
Position Title	Name	Department	Institution
Activity Co-Leader	Dongyan Xu	Computer Science	Purdue University
Activity Co-Leader	Sewoong Oh	Computer Science and Engineering	University of Washington
Faculty Researcher	Radha Poovendran	Electrical Engineering	University of Washington
Graduate Student	Cheng-yun Yang	Computer Science	Purdue University

Project: Robust and Private Transfer Learning Thrust SEC-2: Intelligent Agents for Detection Research/Technology -- Direct Support			
Position Title	Name	Department	Institution
Activity Co-Leader	Sewoong Oh	Computer Science and Engineering	University of Washington
Activity Co-Leader	Elisa Bertino	Computer Science	Purdue University
Postdoctoral Scholar	Maggie Huang	Computer Science	Purdue University
Graduate Student	Saket Gollapudi	Computer Science and Engineering	University of Washington

Project: Bridging the Gap Between Research and Practice in Cyber Threat Intelligence: A Study of CTI Generation, Sharing, and Application Thrust SEC-3: Intelligent Agents for Attribution Research/Technology -- Direct Support			
Position Title	Name	Department	Institution
Activity Co-Leader	Giovanni Vigna	Computer Science	UC Santa Barbara
Activity Co-Leader	David Evans	Computer Science	University of Virginia
Graduate Student	Saastha Vasan	Computer Science	UC Santa Barbara
Graduate Student	Hadjer Benkraouda	Computer Science	UI Urbana-Champaign
Graduate Student	Doguhan Yeke	Computer Science	Purdue University
Graduate Student	Chase Xi Jiang	Computer Science	University of Chicago

Project: Synthetic Generation of High-Fidelity Network Attack Logs Thrust SEC-3: Intelligent Agents for Attribution Research/Technology -- Direct Support			
Position Title	Name	Department	Institution
Activity Co-Leader	Nick Feamster	Computer Science	University of Chicago
Activity Co-Leader	Wenke Lee	Computer Science	Georgia Tech
Graduate Student	Andrew Chu	Computer Science	University of Chicago

Project: A Heterogeneous Malware Classification and Characterization Ensemble Thrust SEC-3: Intelligent Agents for Attribution Research/Technology -- Direct Support			
Position Title	Name	Department	Institution
Activity Co-Leader	Dongyan Xu	Computer Science	Purdue University
Activity Co-Leader	Giovanni Vigna	Computer Science	UC Santa Barbara
Graduate Student	Solomon Sonya	Computer Science	Purdue University
Faculty Researcher	Chris Kruegel	Computer Science	UC Santa Barbara

Project: An adaptable defense for LLM-based computer use agents Thrust SEC-4: Intelligent Agents for Response and Recovery Research/Technology -- Direct Support			
Position Title	Name	Department	Institution
Activity Co-Leader	Wenke Lee	Computer Science	Georgia Tech
Activity Co-Leader	David Wagner	Electrical Engineering and Computer Science	UC Berkeley
Graduate Student	Leyan Pan	Computer Science	Georgia Tech

Globally Accessible Testing Environment (GATE) Research/Technology & Broadening Participation -- Direct Support			
Position Title	Name	Department	Institution
Activity Leader	Berkay Celik	Computer Science	Purdue University
Software Engineer	Kyle Quinlan	Institute Staff	UC Santa Barbara
Activity Leader	Wenke Lee	Computer Science	Georgia Tech
Faculty Researcher	Dongyan Xu	Computer Science	Purdue University
Faculty Researcher	Giovanni Vigna	Computer Science	UC Santa Barbara
Faculty Researcher	Gang Wang	Computer Science	UI Urbana-Champaign
Postdoctoral Scholar	Abdulellah Alsaheel	Computer Science	Purdue University
Faculty Researcher	Bo Li	Computer Science	University of Illinois at Urbana-Champaign
Faculty Researcher	David Evans	Computer Science	University of Virginia
Graduate Student	Doguhan Yeke	Computer Science	Purdue University
Graduate Student	Hongyu Cai	Computer Science	Purdue University
Graduate Student	Jizhou Chen	Computer Science	Georgia Tech
Graduate Student	Yupeng Yang	Computer Science	Georgia Tech
Faculty Researcher	Radha Poovendran	Electrical and Computer Engineering	University of Washington

Affiliated Faculty	Bhaskar Ramasubramanian	Computer Science	Western Washington University
Postdoctoral Scholar	Dinuka Sahabandu	Computer Science	Western Washington University

International Capture the Flag (iCTF) Online Security Training Exercises for High School and Undergraduate Students Education and Workforce Development & Broadening Participation -- Direct Support			
Position Title	Name	Department	Institution
Activity Leader	Giovanni Vigna	Computer Science	UC Santa Barbara
Faculty Researcher	Tonya Fields	Cyber Security	Norfolk State University
Graduate Student	Saastha Vasan	Computer Science	UC Santa Barbara
Software Engineer	Kyle Quinlan	Institute Staff	UC Santa Barbara
Graduate Student	Suraj Suresh	Computer Science	UC Santa Barbara
Graduate Student	Ravindu de Silva	Computer Science	UC Santa Barbara
Graduate Student	Zihao Jerry Su	Computer Science	UC Santa Barbara
Graduate Student	Praneeth Balasubramanian	Computer Science	UC Santa Barbara
Graduate Student	Junmin Jimmy Zhu	Computer Science	UC Santa Barbara
Graduate Student	Suyue Charlie Guo	Computer Science	UC Santa Barbara
Graduate Student	Luca Simei	Computer Science	UC Santa Barbara
Graduate Student	Leonardo Porpora	Computer Science	UC Santa Barbara
Graduate Student	Jesse Cobb	Computer Science	UC Santa Barbara
Graduate Student	Stijn Pletinckx	Computer Science	UC Santa Barbara
Undergraduate Student	Gabriel Pizarro	Computer Science	UC Santa Barbara
Graduate Student	Minjun Elena Long	Computer Science	University of Virginia
Graduate Student	Zhun Wang	Computer Science	UC Berkeley

Graduate Student	Daniel Baumgartner	Computer Science	Rutgers University
Graduate Student	Hadjer Benkraouda	Computer Science	UI Urbana-Champaign
Graduate Student	Saleh Momeni	Computer Science	UI Chicago
Graduate Student	Chase Xi Jiang	Computer Science	University of Chicago
Volunteer Teacher	Jeremy Schwartz	High School Teacher	U Chicago Lab School

Project: zero2hero: Leveraging CTF-based Tools for Cybersecurity and AI Education Education and Workforce Development & Broadening Participation -- Direct Support			
Position Title	Name	Department	Institution
Co Leader	Giovanni Vigna	Computer Science	UC Santa Barbara
Co Leader	Tonya Fields	Cyber Security	Norfolk State University

Project: Composable Curriculum for Cybersecurity in Machine Learning and Generative AI: A Collaborative Educational Initiative Education and Workforce Development & Broadening Participation -- Direct Support			
Position Title	Name	Department	Institution
Co Leader	Radha Poovendran	Electrical and Computer Engineering	University of Washington
Co Leader	Wenke Lee	Computer Science	Georgia Tech
Faculty Researcher	Bhaskar Ramasubramanian	Computer Science	Western Washington University
Faculty Researcher	Giovanni Vigna	Computer Science	UC Santa Barbara

High School Outreach Education and Workforce Development & Broadening Participation-- Direct Support			
Position Title	Name	Department	Institution
Activity Leader	Giovanni Vigna	Computer Science	UC Santa Barbara
Managing Director	Tim Robinson	Institute Staff	UC Santa Barbara
Graduate Student	Stijn Pletinckx	Computer Science	UC Santa Barbara
Undergraduate Student	Gabriel Pizarro	Computer Science	UC Santa Barbara

Summer Research Experience Education and Workforce Development -- Direct Support			
Position Title	Name	Department	Institution
Activity Leader	Tim Robinson	Institute Staff	UC Santa Barbara
Faculty Researcher	Giovanni Vigna	Computer Science	UC Santa Barbara
Faculty Researcher	João Hespanha	Electrical and Computer Engineering	UC Santa Barbara
Faculty Researcher	Jie Gao	Computer Science	Rutgers University
Postdoctoral Scholar	Yigitcan Kaya	Computer Science	UC Santa Barbara
Undergraduate Student Intern	Edward Xiong	Computer Science	Massachusetts Institute of Technology
Undergraduate Student Intern	William Guo	Computer Science	University of Pennsylvania
Graduate Student Intern	Megan Gross	Artificial Intelligence	San Jose State University
Graduate Student Intern	Anoushka Sawant	Artificial Intelligence	San Jose State University
Undergraduate Student Intern	Pedro Gonzalez	Computer Science	Cal State University, San Bernardino
Undergraduate Student Intern	Erik Feng	Computer Science	UC Santa Barbara
Undergraduate Student Intern	Arjun Chopra	Computer Science	Cal Poly SLO
Undergraduate Student Intern	Saqif Ayaan Sudheer	Computer Science	UC Santa Barbara
Undergraduate Student Intern	Brian Lee	Computer Science	University of Chicago

Appendix D: Project Updates (June 2025-May 2026)

Summary slides follow

Thrust AI-1: Learning and Reasoning with Domain Knowledge

Challenge

Current statistical AI techniques lack the capability of integrating domain knowledge and logic reasoning inference ability for the prediction and decision-making tasks.

Current approaches lack scalability.

Overarching Goal

Lay the core theoretical foundation of explaining and reasoning and will remain one of the differentiating themes of ACTION.

Research Threads

- Leverage graph reasoning for proactive cybersecurity defenses, overcoming scalability, data quality, and adversarial challenges.
- Integrate knowledge graphs and LLMs to drive accurate threat analysis and automated cybersecurity defenses.
- Develop lifelong learning systems for continuous anomaly detection and classification to adaptively counter evolving cyber threats.

Planned goals, June 2025 - May 2026

- Develop time-dependent dynamic graph models to capture evolving cyber threats and enable proactive, real-time threat intelligence and intervention.
- Tune LLMs to reason over cybersecurity time series data, enabling automated analysis of complex temporal patterns in cyber environments.
- Advance lifelong learning algorithms that autonomously discover and adapt to new threats without relying on fixed datasets or human-labeled training data.

Achievements this period

- Developed new dynamic graph methods built on counterfactual reasoning, information bottleneck, and persistent homology, contributing to multiple manuscripts.
- Produced a paper accepted to ACL 2026 on inferring events from time series using language models, with next phase moving toward cybersecurity threat detection applications.
- Advanced autonomous lifelong learning algorithms that detect and incorporate new threats without human-labeled data
- Developed unlearning methods to correct erroneous knowledge post-deployment.

Ambuj Singh[1,2], Jie Gao[1,3], David Evans[1,4], Sewoong Oh[1,5], Xiaoming Huo[1,6], Wenke Lee[1,7], Bing Liu[1,8]
[1] Faculty, [2] UC Santa Barbara, [3] Rutgers, Computer Science, [4] University of Virginia, Computer Science, [5] University of Washington, Computer Science and Engineering, [6] Georgia Tech, Statistics and Data Science, [7] Georgia Tech, Computer Science, [8] University of Illinois, Chicago, Computer Science

Thrust AI-2: Human-Agent Interaction

Challenge

Current human-AI systems lack principled methods for supporting human responses, soliciting human intelligence at scale, and enabling effective collaboration in mixed human-agent teams.

Overarching Goal

Develop principled methods across decision-making, learning, and teaming paradigms to optimize human-agent interaction.

Research Threads

- Improve human-agent decision-making by promoting calibrated trust, modeling human behavior, and designing agents that adapt to human responses.
- Augment autonomous agents by scalably soliciting human domain knowledge and intuitive feedback to guide agent learning.
- Optimize human-agent teaming through principled task allocation and effective communication strategies for interdependent collaboration.

Planned goals, June 2025 - May 2026

- Design adaptive AI assistance that personalizes presentation formats and content to individual decision makers, optimizing multiple objectives simultaneously.
- Develop LLM-powered agents that enhance human-agent collective decision-making by improving team memory, information aggregation, and fairness in high-stakes cybersecurity contexts.

Achievements this period

- Developed an adaptive AI ensemble that switches between trust-building and performance-boosting models, with a follow-on framework using counterfactual analysis to reduce over-reliance; resulting in publications at AAAI 2026 and CHI 2026 (honorable mention).
- Designed experiments on LLM-facilitated team decision-making, including transactive memory building and hidden-profile information aggregation, with findings accepted at IC2S2 2026.

Ming Yin[1,2], Ambuj Singh [1,3]

[1] Faculty, [2] Purdue University, Computer Science, [3] UC Santa Barbara, Computer Science

Thrust AI-3: Multi-Agent Collaboration

Challenge

Current multi-agent frameworks lack principled methods for modeling network capabilities, handling agent heterogeneity and asynchrony, and accounting for adversarial agents and conflicting worldviews.

Overarching Goal

Develop a comprehensive multi-agent collaboration framework spanning network modeling, reasoning, and adversary-aware learning to enable robust and effective agent collaboration.

Research Threads

- Model multi-agent networks to enable resilient, trust-aware knowledge sharing across distributed agents.
- Enable agents to efficiently query relevant knowledge and reason robustly under uncertainty and missing information.
- Train multi-agent systems to learn policies resilient to adaptive adversarial attacks and heterogeneous threat vectors.
- Secure LLM-powered agents through controlled data access, malicious input filtering, and enforced protection objectives.

Planned goals, June 2025 - May 2026

- Develop techniques that enable agents to securely control information disclosure and make robust decisions when receiving potentially corrupted information from malicious agents.
- Build a secure multi-agent healthcare system that orchestrates specialized AI agents to reliably integrate clinical knowledge, follow safety policies, and coordinate complex medical workflows.

Achievements this period

- Developed techniques for enforcing security policy on LLM agent outputs and enabling agents to reason under potentially corrupted information
  - Two academic submissions and one manuscript in preparation.
- Introduced "Green Shielding," a user-centric framework for characterizing how routine input variations shift LLM behavior
  - Paper posted to arXiv.

Jie Gao [1,2], Ninghui Li [1,3], Bo Li [1,4], Bin Yu [1,5]

[1] Faculty, [2] Rutgers, Computer Science, [3] Purdue, Computer Science, [4] University of Illinois, Urbana-Champaign, Computer Science, [5] UC Berkeley, Statistics

Thrust AI-4: Strategic Gaming and Tactical Planning

Challenge

Current game-theoretic approaches lack the ability to reason about dynamic attacker behavior, provide worst-case performance guarantees, or handle fast, high-volume information flows.

Overarching Goal

Establish a composable defensive layer that enables intelligent agents to make robust, adaptive decisions under uncertainty through advances in strategy games, threat modeling, and dynamic tactical planning.

Research Threads

- Develop non-equilibrium cyber-AI game models to compute best agent responses at each stage of the cyber kill-chain.
- Model adversary behavior using probabilistic and Markov-based approaches to quantify attack risks and anticipate stealthy threats.
- Plan agent tactics through deception, moving target defense, and privacy-preserving obfuscation to guarantee robust performance.

Planned goals, June 2025 - May 2026

- Develop no-regret learning algorithms that adapt to dynamic, multi-step attack scenarios and rare events across detection, attribution, and response applications.
- Deploy a Dynamic Profiling Agent that uses LLM-guided Monte Carlo simulation to identify high-risk attack paths and focus vulnerability analysis on the most critical code regions.
- Reveal limitations of ML-based APT detection and develop new strategies for modeling and evading low-and-slow attack detection.

Achievements this period

- Developed region-aware learning algorithms with optimal regret bounds that adapt to evolving adversarial threats, extended to defend cyber-physical systems against stealthy attacks.
- Showed that opponent-informed exploration protects Q-learning policies against adversarial deception in large-state-space RL problems.
- Built "Klleem," an LLM-guided Dynamic Profiling Agent that outperforms state-of-the-art vulnerability discovery.
- Initiated a model-based study of low-and-slow APT evasion, targeting underexplored limitations of ML-based audit log detection.

Radha Poovendran [1,2], João Hespanha [1,3], Wenke Lee [1,4], Christopher Kruegel [1,5], Dongyan Xu [1,6], Sewoong Oh [1,7]

[1] Faculty, [2] University of Washington, Electrical and Computer Engineering, [3] UC Santa Barbara, Electrical and Computer Engineering, [4] Georgia Institute of Technology, Computer Science, [5] UC Santa Barbara, Computer Science, [6] Purdue University, Computer Science, [7] University of Washington, Computer Science and Engineering

AI Institute for Agent-based Cyber Threat Intelligence and Operation

Globally Accessible Testing Environment (GATE)

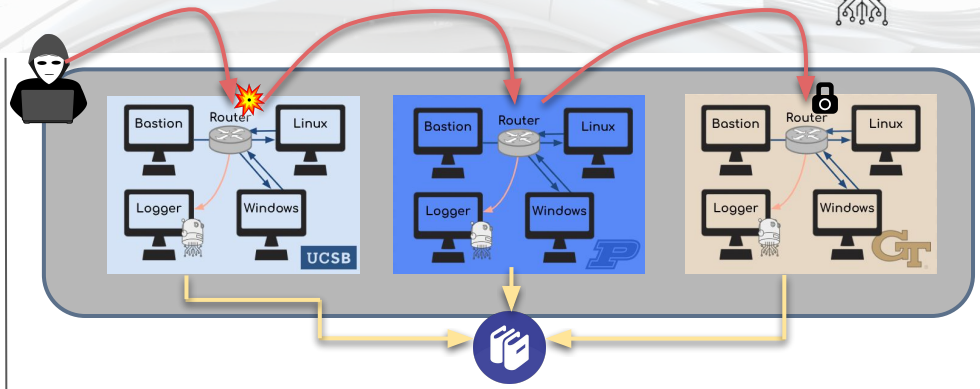
Overarching Goal

Develop a virtual, integrated testbed for ACTION, enabling

- User-friendly experiment set up
- Always up-to-date configuration
- Remote access through web-based APIs
- Flexible, extensible scenarios, based in real world

Challenges

- Scalability requires robust technology solutions
- Network architecture is complicated; should facilitate seamless interaction between agents and modeled humans.



Demonstrating GATE's potential: UCSB, Purdue, and Georgia Tech created networks using the GATE-server communication channel to showcase deployed agent prototypes creating and sharing knowledge in response to a simulated attack.

Proposed Solutions

- Incremental development that is prototype-centric
- Engagement with cybersecurity community (with a vested interest in agent-based solutions) to identify broad features/requirements
- Ability to showcase reproducible, high fidelity artifacts from a real-world attack on the cloud-based testbed
- Input from ACTION performers across all research thrusts and activities on required databases, libraries, documents, etc.

Achievements this period

- Developed a multi-agent framework that automatically identifies adversarial failure scenarios in embodied robotic AI systems.
- Built an attack agent that bypasses state-of-the-art safety defenses in embodied AI, including Asimov-style rule enforcement.
- Created benchmark agents that systematically evaluate attack/defense agents and measure robots' ability to correctly refrain from action.
- Developed skill-based and Codex agents to autonomously solve reverse engineering tasks without human intervention.
- Two manuscripts under preparation for submission.

Z. Berkay Celik[1,2], Dongyan Xu[1,2], Wenke Lee[1,3], Giovanni Vigna [1,4] et al.

[1] Faculty, [2] Purdue University, Computer Science, [3] Georgia Institute of Technology, Computer Science, [4] UC Santa Barbara, Computer Science

Thrust SEC-1: Intelligent Agents for Threat/Vulnerability Assessment

Challenge

Traditional vulnerability analysis techniques struggle to scale, generate excessive false positives, and provide limited coverage in increasingly complex modern software and hardware systems.

Overarching Goal

Develop an AI-agent-powered framework integrating machine learning and LLMs to improve the efficiency and effectiveness of vulnerability discovery, understanding, and remediation across both individual components and composite systems.

Research Threads

- Use AI agents to identify high-risk code regions and orchestrate scalable, targeted vulnerability analysis.
- Develop an LLM-powered patching system that reasons about vulnerability root causes and verifies secure fixes.
- Assess and harden AI agents and LLMs against adversarial vulnerabilities such as prompt injection and malicious fine-tuning.
- Extend vulnerability assessment to cyber-physical systems via AI-driven digital twin modeling, simulation, and fuzzing.

Planned goals, June 2025 - May 2026

- Build an LLM-powered reverse engineering agent that automates end-to-end binary analysis through hypothesis generation and verification, while systematically identifying and improving upon LLM weaknesses in the RE lifecycle.
- Deploy a Dynamic Profiling Agent that uses LLM-guided Monte Carlo simulation to identify high-risk attack paths and focus vulnerability analysis on the most critical code regions.

Achievements this period

- Expanded evaluation to 70+ CTF reverse-engineering challenges, introducing a systematic rubric to identify LLM failure points across the RE lifecycle.
- Discovered that leading LLMs are susceptible to cognitive biases, including halo, framing, and anchoring effects, in code vulnerability detection; paper in preparation.
- Built "Klleem," a Dynamic Profiling Agent combining LLM-guided seed generation with fuzzing and symbolic execution, outperforming state-of-the-art vulnerability discovery.

Gang Wang [1,2], Berkay Celik [1,3], Wenke Lee [1,4], Christopher Kruegel [1,5]

[1] Faculty, [2] University of Illinois, Urbana-Champaign, Computer Science, [3] Purdue University, Computer Science, [4] Georgia Tech, Computer Science, [5] UC Santa Barbara, Computer Science,

Thrust SEC-2: Intelligent Agents for Detection

Challenge

Current detection approaches are brittle against incomplete data, adaptive adversaries, and uncertain contexts, producing unreliable and difficult-to-interpret models.

Overarching Goal

Develop robust, interpretable detection agents that handle threat evolution, evasion, and uncertainty through advances in training, multi-agent collaboration, and privacy-preserving data sharing.

Planned goals, June 2025 - May 2026

- Develop a context-adaptive ensemble malware classifier with diverse detectors and customizable combination policies to reduce false positives across varied contexts.
- Build a unified guardrail framework that uses safety reasoning and dynamic knowledge integration to protect agentic systems against adversarial threats.
- Develop RL-based intrusion response policies that automatically optimize the trade-off between containing network attacks and minimizing false alert disruption.
- Reveal ML limitations in audit log-based APT detection and develop new gaming strategies for modeling and evading low-and-slow attacks.
- Develop transfer learning algorithms that remain robust against combined evasion, poisoning, and private data extraction attacks across source and target domains.

Research Threads

- Train robust detection agents that adapt to concept drift and evasion using transfer learning and LLM-based reasoning.
- Automate reproduction and synthesis of real-world multi-step attack chains for high-quality threat detection training data.
- Enable secure cross-organization threat data sharing through privacy-preserving, poisoning-robust training methods.
- Deploy heterogeneous detection and deception agents that collaborate to detect lateral movement and stealthy attacks.

Achievements this period

- Developed MAPD, a post-deployment adaptation method that corrects for policy discrepancy using minimal feedback.
- Built a unified guardrail framework validated across diverse agentic systems.
- Developed a dual-agent privilege-separation framework to defend against prompt injection attacks.
- Showed state-of-the-art intrusion detection systems can be systematically evaded; paper submitted.

Christopher Kruegel [1,2], David Evans [1,3], David Wagner [1,4], Dawn Song [1,4], Bo Li [1,5], Gang Wang [1,5], Dongyan Xu [1,6], Radha Poovendran [1,7], Sewoong Oh [1,8], Elisa Bertino [1,6]
[1] Faculty, [2] UC Santa Barbara, Computer Science, [3] University of Virginia, Computer Science, [4] UC Berkeley, Computer Science, [5] University of Illinois, Urbana-Champaign, Computer Science, [6] Purdue University, Computer Science, [7] University of Washington, Electrical Engineering, [8] University of Washington, Computer Science and Engineering

Thrust SEC-3: Intelligent Agents for Attribution

Challenge

Current CTI workflows are manual, error-prone, and slow, hindered by disparate data formats, ineffective sharing, and AI techniques insufficient for precise, real-time threat classification.

Overarching Goal

Develop an AI-driven CTI system that automates threat extraction, standardizes intelligence sharing, and enables real-time detection and classification of cyber threats across organizations.

Research Threads

- Automate CTI extraction from raw logs using LLMs and graph-based AI to identify threat indicators and correlate attack patterns.
- Automate CTI codification and sharing using NLP and privacy-preserving federated learning across organizations.
- Automate CTI application to network security through LLM-driven threat hunting and reinforcement learning-based defense adjustments.
- Develop an automated malware classification framework using NLP ensemble models to improve family attribution and characterization.

Planned goals, June 2025 - May 2026

- Systematically compare academic CTI techniques with real-world practitioner workflows to identify gaps and opportunities for better-aligned AI-driven CTI research.
- Generate synthetic, high-fidelity network attack logs using diffusion-based models to enable scalable training and benchmarking of CTI extraction models in data-scarce environments.
- Develop a live malware dataset and heterogeneous classification ensemble to improve generalization and accuracy in malware family attribution and behavior characterization.

Achievements this period

- Reviewed 127 papers and surveyed 18 practitioners to identify key CTI automation bottlenecks; paper submitted.
- Extended NetDiffusion and NetSSM to generate high-fidelity, multi-stage synthetic attack traces; published at SIGMETRICS 2024 (NetDiffusion) and to appear at CoNEXT 2026 (NetSSM).
- Enriched MABEL with 1,000+ attributes per sample; published heterogeneous ensemble classifier at ESORICS SECAI 2025.

Giovanni Vigna [1,2], David Evans [1,3], Nick Feamster [1,4], Wenke Lee [1,5], Dongyan Xu [1,6]

[1] Faculty, [2] UC Santa Barbara, Computer Science, [3] University of Virginia, Computer Science, [4] University of Chicago, Computer Science, [5] Georgia Tech, Computer Science, [6] Purdue University, Computer Science

Thrust SEC-4: Intelligent Agents for Response and Recovery

Challenge

Current response and recovery approaches are ineffective against dynamic multi-step attacks, vulnerable to adversarial manipulation, and too ad-hoc to generalize across attack scenarios.

Overarching Goal

Develop a cooperative, AI-driven response and recovery framework that leverages intelligent agents to devise and execute adaptive, generalizable strategies against complex, evolving cyber attacks.

Research Threads

- Build an autonomous attack management system using AI-driven reasoning and dynamic tools to detect and contain multi-step attacks.
- Enable distributed situational awareness through collaborative multi-agent reasoning across hosts and network segments.
- Train response agents via reinforcement learning in simulated cyber ranges to develop adaptive response and hardening policies.
- Harden response agents through secure LLM controls, automated code generation, and explainable decision-making.

Planned goals, June 2025 - May 2026

- Develop a defense system for LLM-based computer use agents that learns security rules through human feedback to detect and prevent prompt injection and other adversarial attacks at test time.
- Evolve GATE into a flexible, scalable testbed by improving network configurability, deploying ACTION research agents, and connecting to external testbeds for high-fidelity experimentation.

Achievements this period

- Began developing ‘Secure ReAct,’ a prompt-injection-resistant agent design pattern, and exploring its extension to vision-language models.
- Built multi-agent framework identifying robotic AI failure scenarios; showed state-of-the-art safety defenses can be bypassed.
- Advanced GATE's reverse engineering capabilities through autonomous, skill-based RE agents.

Wenke Lee [1,2], David Wagner [1,3], Z. Berkay Celik [1,4]

[1] Faculty, [2] Georgia Institute of Technology, Computer Science, [3] UC Berkeley, Computer Science, [4] Purdue University, Computer Science

Education: Using Capture the Flag for Security + AI Training

Project Overview

- Capture the Flag security exercises are fun, approachable
- Jeopardy-style CTF incorporating AI in three modes:
  - AI as the subject of security challenges
  - Human-agent teaming
  - Human-led, agent+agent teaming (Novel)
- Separate tracks for high school and undergraduate
- Year 3 goals: grow iCTF, sponsored by ACTION
- iCTF can be re-packaged for education and workforce development (potential for industry collaboration)



CTF camp participants at NSU 2025



2026 iCTF flier

Challenges

- Many HS and college students are unprepared for the demands of CTF, or they are hesitant to participate
- Advertising the event efficiently
- Scaling the infrastructure while ensuring participant integrity

Solutions

- Increased outreach in schools; boots on the ground approach
- Online videos/tutorials to help students through challenges asynchronously
- Efficient and novel challenge design for different levels of students

Achievements this period

- NSU launched CTF Summer Camps for HS, College, and Veterans
- Surveys showed participants overwhelmingly believed the experience to be beneficial
- Agentish platform launched first-ever “Agentic CTF”
  - “Scratch” for agentic programming
  - Teaches development and deployment of AI agents
- New walkthroughs planned for design in 2026 to prepare for next round of CTF

Saastha Vasan [1,3], Giovanni Vigna [2,3], Tonya Fields [2,4]

[1] Graduate Student, [2] Faculty, [3] UC Santa Barbara, Computer Science, [4] Norfolk State University, Cybersecurity

Education: zero2hero: Leveraging CTF-based Tools for Cybersecurity and AI Education

Challenge

The cybersecurity workforce faces a critical national security gap as sophisticated state-sponsored attacks intensify and AI becomes increasingly central to security solutions, yet accessible training pathways remain limited.

Overarching Goal

Develop an educational program that teaches cybersecurity and AI concepts from the ground up, leveraging CTF-based learning to build workforce capability without requiring prior knowledge.

Research Threads

- Develop a fully online, scalable CTF-based educational platform teaching cybersecurity and AI from the ground up using the pwn.college platform, requiring no prior knowledge.
- Build challenge-specific LLM-based AI tutors for each CTF exercise, going beyond the basic AI tutor currently available in pwn.college.
- Develop a framework using logic-based declarative approaches to formally verify that students have internalized and can transfer specific security and AI skills across different contexts.

Planned goals, June 2025 - May 2026

- Curriculum covering basic cryptography, network security, application security, web security, and core AI concepts each paired with CTF exercises.
- Challenge-specific AI tutoring system leveraging knowledge representation (Thrust AI-1) and human-agent interaction research (Thrust AI-2).
- Pilot deployment in experimental classes at UCSB and NSU.
- First generation of an interactive online educational experience combining multi-modal learning, AI-based tutoring, and skill acquisition verification.

Achievements this period

- NSU Walkthroughs of picoCTF and SANS CTF to supplement student learning.
- Two modular courses on vulnerability analysis at UCSB
  - Conceptual course covering security vulnerabilities at multiple levels of abstraction
  - Hands-on course integrating CTF challenges and pwn.college modules.
- Introduced 40 new CTF challenges, each accompanied by walkthroughs and AI prompts to support agent-assisted student learning.

Giovanni Vigna [1,2], Tonya Fields [1,3]

[1] Faculty, [2] UC Santa Barbara, Computer Science, [3] Norfolk State University, Cybersecurity

Education: Composable Curriculum for Cybersecurity in Machine Learning and Generative AI: A Collaborative Educational Initiative

Challenge

- Cybersecurity is rapidly transforming as ML and generative AI are integrated into critical infrastructure.
- Threats now span both traditional attack vectors and AI-specific vulnerabilities.
- Existing curricula are often static and siloed, failing to reflect these evolving realities.

Overarching Goal

Develop a flexible, modular, and hands-on curriculum that covers these advanced concepts at the intersection of AI and Cybersecurity and promotes deep understanding through self-guided, inquiry-based learning supported by AI tutors.

Planned goals, June 2025 - May 2026

- Launch an AI-Tutor prototype capable of guiding students through examples of varying difficulty and generating adaptive quiz variations.
- Develop and pilot week-long modular course units that can be offered as standalone mini-courses or combined into full 10-week courses at participating institutions and NSF summer programs.
- Make all course modules, lab materials, and datasets publicly available via GitHub, with video lecture options and online office hours to support scalable, asynchronous delivery.

Research Threads

- Develop a modular, hands-on curriculum spanning "Security for GenAI" and "AI for Cybersecurity," designed to serve undergraduate, graduate, and working professional audiences.
- Offer six composable course modules — including ML for Cybersecurity, AI Alignment, adversarial robustness, and edge intelligence — each pairing lectures with hands-on labs grounded in ACTION research.
- Deploy an AI-Tutor using the Socratic method to foster critical thinking, supplemented by Jupyter notebooks, asynchronous video lectures, and GitHub-hosted materials for broad accessibility.

Achievements this period

- Teacher Chatbot, Lowkey Genius deployed on Render
- Draft of curriculum for AVATAR: novel course on genAI for vulnerability analysis
- Connections w/ NSF and NSA Centers of Academic Excellence for Cyber Security
  - Machine Learning for Cybersecurity (ML4Cyber)
  - AI Alignment for Safe Downstream Applications
  - AI for Edge Intelligence (TinyML)
  - AI/ML based Security Analytics
  - Robustness of AI to Adversarial Manipulation
  - Iterative Approaches AI/ML Security

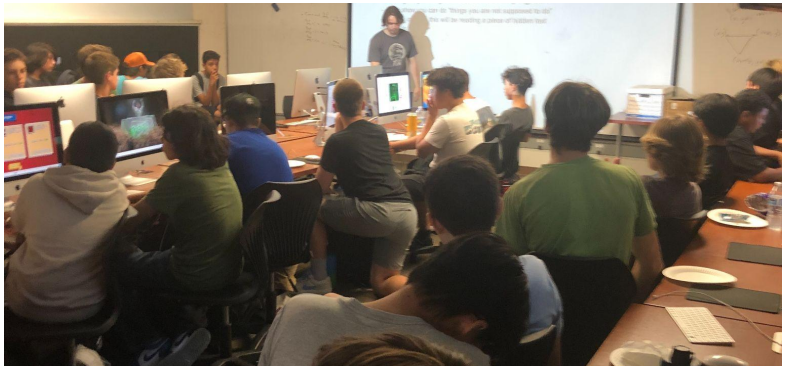
Radha Poovendran [1,2], Wenke Lee [1,3], Bhaskar Ramasubramanian [1,4], Giovanni Vigna [1,5]

[1] Faculty, [2] University of Washington, Electrical and Computer Engineering, [3] Georgia Tech, Computer Science, [4] Western Washington University, Computer Science, [5] UC Santa Barbara, Computer Science

Outreach: Empowering High School Students with Computer Science and Cyber Security Education

Project Overview

- Goal: Increase participation in CS at the high school level through the introduction of cybersecurity training
- In-person outreach, in collaboration with high school instructors
- Targets freshmen and sophomores in high school, especially from traditionally underrepresented populations
- If institutionalized as a student-led club, online courses generated from iCTF data can be used for curriculum
- Project integrates well with iCTF and can be instantiated at other high schools near ACTION campuses and beyond



Pictured: Stijn Pletinckx/ACTION Outreach at SB High School.

Challenges

- Engaging the students, who are already busy, or might be intimidated by CS concepts
- Finding/training students to lead clubs (not all students are leaders)
- Supporting graduate and undergraduate tutors to work in high schools
- Setting aside time at HS for club activities

Solutions

- Secure funding to support dedicated graduate students
- Train teachers in problem solving curriculum
- Empower the HS students with the ability to teach
- Clubs meet at lunch, or after school

Achievements this period

- Launch of student-led hacking club at SB High School
- Integration with CS teacher
- Bi-weekly sessions led by PhD student
- New hacking team with participation in iCTF and PicoCTF
- Shareable curriculum, with pointers to PicoCTF material
- Presentation at HS in Urbana, Illinois
- New connections between student club in Chicago with research lab at U Chicago
- Preparations for summer camp targeting HS in Norfolk, VA
- Teacher training held at UCSB, June 2025

Stijn Pletinckx [2,5], Ilya Grishchenko [4,5], Laboni Sarker [2,5], Tim Robinson [3,5], Giovanni Vigna [1,5],

[1] Faculty, [2] Graduate Student, [3] Program Manager, [4] Postdoctoral Scholar, [5] UC Santa Barbara, Computer Science

Outreach: Summer Research Experience for Undergraduate and Graduate Students

Project Overview

- 8-week summer research experience, directed toward students who typically do not have access to research labs
- Opportunities for directed research, professional development, and networking for undergrads and grads
- Promotion at ACTION campuses, but primarily at MSIs
- Focus is on broadening participation and providing new research experiences



Challenges

- Advertising to obtain a large applicant pool
- Preparing appropriate, mini-research projects
- Finding available mentors for participants

Solutions

- Establish recruiting partnerships with student groups/organizations across ACTION
- Mentor training through collaboration with established campus programs on ACTION campuses

Achievements this period

- 8 undergraduates recruited. 2 graduate students + 1 high school student
- Multiple academic paper submissions from results;
  - NeurIPS 2025 workshop publication
  - AAMAS'26 conference publication
  - CSUSB Undergraduate Symposium (Best Poster Award)
- Teams of 2-4 alongside senior-level ACTION graduate student, postdoc, or faculty
- Help from ACTION Student Advisory Council to select students
- Projects at UCSB, Rutgers in 2025
- Cohort for 2026, ready to start in May/June at UCSB, Rutgers, UW, and Georgia Tech

Tim Robinson [1,5], Giovanni Vigna [2,5], Yigitcan Kaya [4,5], Jie Gao [2,6], ACTION Graduate Student Advisory Council [3,4,7]

[1] Program Manager, [2] Faculty, [3] Graduate Student, [4] Postdoctoral Scholar, [5] UC Santa Barbara, Computer Science, [6] Rutgers, Computer Science, [7] All ACTION Campuses.