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 2 Annual Report
June 2024 - May 2025

Approved for Public Release

NSF Cooperative agreement: 2229876

Table of Contents

Project Summary **4**

Participant Tables **5**

TABLE 1 – PRINCIPAL ORGANIZATIONS OF THE INSTITUTE 5

TABLE 2 - LIST OF THE LEADERSHIP TEAM 6

TABLE 3 - LIST OF MAJOR INSTITUTE ACTIVITY PARTICIPANTS 9

TABLE 4 - LIST OF OTHER COLLABORATORS AND PARTNERS 12

TABLE 5 - ADVISORY BOARDS 12

Executive Summary **13**

Narrative: **16**

1.0 Vision 16

2.0 Strategic Plan 17

3.0 Essential Functions of the Institute 22

3.1 Research 22

3.1.1 Foundational AI research 23

Thrust AI-1: Learning and Reasoning with Domain Knowledge 23

Thrust AI-2: Human-Agent Interaction 24

Thrust AI-3: Multi-Agent Collaboration 25

Thrust AI-4: Strategic Gaming and Tactical Planning 25

Globally Accessible Testing Environment (GATE) 26

3.1.2 Use Inspired Research 27

Thrust SEC-1: Intelligent Agents for Threat/Vulnerability Assessment 27

Thrust SEC-2: Intelligent Agents for Detection 28

Thrust SEC-3: Intelligent Agents for Attribution (and Threat Intelligence) 28

Thrust SEC-4: Intelligent Agents for Response and Recovery 29

3.1.3 Benefits of use-inspired research 29

3.2 Nurture and grow the next generation of talent 30

3.2.1. Pre-college education 30

3.2.2 University, College, and Community College Education 31

3.2.3 Workforce development 34

3.2.4 Informal Learning and Education 35

3.3 Broadening participation 35

3.3.1 International Capture the Flag 35

3.3.2 Outreach in High Schools 36

3.3.3 Summer Research Experiences 37

3.4 Multidisciplinary integration and community building 40

3.4.1 Seminars 40

3.4.2 Campus Visits for Community Building 41

3.4.3 Other Meetings and Communication 41
3.5 Multi-organizational synergies/achievements 41
3.6 Knowledge transfer 42
3.6.1 Publications 42
3.6.2 Presentations 43
3.6.3 Other Meetings Related to Knowledge Sharing 44
3.6.4 Other Knowledge Sharing 45
3.7 Impact of the Institute as a nexus point for collaborative efforts 46
3.7.1 Collaborative grants 46
3.7.2 Other collaborative efforts 46
4.0 Evaluation and Assessment 47
4.1. Education and Outreach Evaluation 48
4.2. Response to External Evaluation 49
5.0 Infrastructure and Management 50
5.1. Configuration and Leadership 50
5.1.1 Organizational Chart 50
5.1.2 Ethics Training 52
5.2 Management Systems and Reporting 52
5.2.1 Management Systems 52
5.2.1.1 Role of advisory boards. 52
5.2.1.2 Project management 52
5.2.3 Conflicts of Interest 54
References **55**
Appendix A: Glossary and Acronyms 62
Appendix B: Highlights of Significant Achievements and Impact 65
Highlight 1: Enabling Asymptotic Truth Learning in a Social Network 65
Outcome/Accomplishment: 65
Key Achievements: 65
Real-world Impacts: 65
Relevant ACTION Research Thrust/Thread: 66
Reference: 66
Highlight 2: Learning with contextual information in non-stationary environments 66
Outcome/Accomplishment: 66
Key Achievements: 66
Practical Examples: 67
Real-world Impacts: 67
Relevant ACTION Research Thrust/Thread: 67
Reference: 67
Highlight 3: Toward Data-driven Holistic Security Posture Analysis using AI-generated Attack Graphs 67

Outcome/Accomplishment: 67
Key Achievements: 67
Real-world Impacts: 68
Relevant ACTION Thrust/Thread: 68
Reference: 68
Highlight 4: ArtPrompt: ASCII Art-based Jailbreak Attacks against Aligned LLMs 69
Outcome/ Accomplishment: 69
Key Achievements: 69
Real-World Impact: 69
Relevant ACTION Thrust/Thread: 70
Reference: 70
Highlight 5: DeepCapa: Identifying Malicious Capabilities in Windows Malware 70
Outcome/ Accomplishment: 70
Real-World Impact: 70
Relevant ACTION Thrust/Thread: 71
Reference: 71
Highlight 6: MABEL and Heterogeneous Ensemble: Advancing Malware Family Attribution with Scalable AI Models and Granular Analysis 72
Outcome / Accomplishment 72
Impact and Benefits 72
Relevant ACTION Thrust/Thread: 73
Reference: 73
Appendix C: List of Institute Projects 74
Appendix D: Project Updates 84

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

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	Christopher 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 Chicago / University of Illinois 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
Managing Director	Tim Robinson	Institute Staff	University of California, Santa Barbara

TABLE 3 - LIST OF MAJOR INSTITUTE ACTIVITY PARTICIPANTS

Thrust AI-1: Learning and Reasoning with Domain Knowledge			
Position Title	Name	Department	Institution
Activity Leader	Ambuj Singh	Computer Science	University of California, Santa Barbara
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	Bo Li	Computer Science	University of Chicago / University of Illinois Urbana-Champaign
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	Xiaoming Huo	Data Science and Statistics	Georgia Institute of Technology
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	João Hespanha	Electrical and Computer Engineering	University of California, Santa Barbara
Faculty Researcher	Radha Poovendran	Electrical and Computer Engineering	University of Washington
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	Christopher Kruegel	Computer Science	University of California, Santa Barbara
Faculty Researcher	Bo Li	Computer Science	University of Chicago / University of Illinois Urbana-Champaign
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	Christopher 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 Test Environment			
Position Title	Name	Department	Institution

Activity Leader	Berkay Celik	Computer Science	Purdue University
Software Engineer	Noah Spahn	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 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

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
Student Advisory Board			

Name	Title	Organization (Department or Division)	Institution or Firm
Shinan Liu (Council Head)	PhD Student	Computer Science	University of Chicago
Fengqing Jiang	PhD Student	Electrical and Computer Engineering	University of Washington
Yujin Potter	Postdoctoral Researcher	Electrical Engineering and Computer Science	University of California, Berkeley
Saastha Vasan	PhD Student	Computer Science	University of California, Santa Barbara
Leyan Pan	PhD Student	Computer Science	Georgia Institute of Technology
Raymond Wijaya Muller	PhD Student	Computer Science	Purdue University
Kevin Lu	PhD Student	Mathematics	Rutgers University
Hadjer Benkraouda	PhD Student	Computer Science	University of Illinois at Urbana-Champaign
Michael Jerge	PhD Student	Computer Science	University of Virginia
Guy Munn	MS Student	Computer Science	Norfolk State University

Executive Summary

Narrative

The ACTION AI Institute's second year of operation saw continued progress along all three of the Institute's driving goals. By generating more than 60 publications that furthered the development of security solutions based on intelligent agents, researchers brought the total count of ACTION publications to 100. ACTION researchers also made significant progress in the development of their Globally Accessible Testing Environment (GATE) and released early versions of their proposed AI-stack.

Regarding education and outreach, the continued hosting of the International Capture the Flag (iCTF) introduced a new cohort of high school and undergraduate hackers to the world of computer security. New classes by ACTION faculty introduced college students and the general workforce to agentic programming and LLMs. The launch of a multi-campus summer internship program in 2024 brought students to AI and computer security labs who might have otherwise missed out on the opportunity.

ACTION brought hands-on computer science training to nearby high schools via the creation of student-led hacking clubs.

Knowledge transfer is also on-going. Complementing the above-referenced list of 100 manuscripts are a series of code and database releases for the general community of AI + Security researchers through GitHub and NAIRR. ACTION faculty have participated in multiple invited presentations, including panels to government agencies and industry.

Most importantly, the Institute's management, with feedback from other ACTION faculty, the Institute's External Advisory Board, and the National Science Foundation have drafted a revision of ACTION's Strategic & Implementation Plan (S&IP) meant to bring the entire community into a more-collaborative and focused Year 3. We note the significant updates made to the S&IP and, where appropriate, take time to describe how operations will be handled differently in the future. It is important for readers to consider that most of ACTION's Year 2 activities follow the format established in the first year of operation.

The report that follows, along with its multiple appendices, includes details on all of ACTION's activities from June 2024 through May 2025, with hints at activities leading up to the Year 2 Annual Review. The report ends with a discussion of management operations, including changes to the management team.

Achievements in the Reporting Period:

i. Foundational AI research
    a. Development of framework for enhancing collective decision making among networks of agents
    b. Method to help agents continue learning and improving, even when faced with unpredictability
    c. Cybersecurity tool for AI-generated attack graphs offers intelligent, automated solution for comprehensive security assessment
ii. Other Institute research
    a. Collaborative project on art-based jailbreak attacks allows focus on agents' reasoning about vulnerabilities
    b. Improvements to a novel methodology for uncovering malicious capabilities in Windows malware

c. Novel paradigm of building and evaluating heterogeneous classification ensembles that can be applied to different domains.

iii. Benefits of use-inspired research
   a. UIUC based team participating in Amazon's Trusted AI Challenge
   b. UCSB based team participating in DARPA AI Cyber Challenge

iv. University and college education
   a. Successful iCTF for undergraduate students
   b. New undergraduate and graduate courses

v. Pre-college education
   a. 60% growth of iCTF for high school students
   b. 12 walkthroughs for introduction to CTF on YouTube
   c. Increased high school outreach beyond Santa Barbara

vi. Workforce development
   a. LLM MOOC

vii. Broadening participation and diversity
   a. Summer internships at UCSB, Rutgers, and Norfolk
   b. Use of evidence-based methodology in outreach
   c. In-person training for high school teachers planned for June 2025

viii. Multidisciplinary integration and community building
   a. ACTION Seminar Series
   b. Institute-wide research group meetings

ix. Multi-organizational synergies/achievements
   a. ACTION faculty co-organized a session on Statistical Methods in Cybersecurity at the Joint Statistical Meeting (JSM 2024)
   b. All-hands-on-deck call for iCTF challenge creation leads to participation from across ACTION Institute
   c. Drafted manuscript: SoK on AI for threat intelligence

x. Knowledge transfer
   a. Multiple invited ACTION faculty presentations on research
   b. Panel participation ACSAC 2024, CARTA 2025, & Billington Cybersecurity
   c. Release of MAGPIE datasets, AI-stack on GitHub, and MABEL to NAIRR

xi. Impact of the institute as a nexus point for collaborative efforts
   a. Collaborative grant on security in AI involves two ACTION Co-PIs and expands reach to a third, non-ACTION professor at another institution
   b. Invitation to Summit for Digital Trust at C4DT in Glion, Switzerland
   c. Multiple opportunities for foreign and domestic collaboration

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. The Institute will also foster collaboration between its member organizations and strategic partners 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

The ACTION Institute's Strategic Implementation Plan (S&IP) guiding our second year of operation is available upon request. We drafted this version of the S&IP (v2) in May 2024 and submitted the document with our first annual report. Our S&IPv2 defines the managerial structure of the ACTION Institute from program inception through May 2025 (first two years). It also outlines the eight research thrusts directing our efforts during that time frame. Each thrust first identifies relevant challenges and then suggests plans for overcoming those challenges. As part of its plan, each thrust is then divided into three interdependent research "threads," and then further divided into various phases, meant to help ACTION researchers track progress toward deliverables (e.g. an AI stack). At times, phases overlap among threads due to the integrated nature of the research. S&IPv2 also describes, at a high level, the various education and outreach activities the Institute would employ to make an impact in the community and future workforce during the first and second years.

In February 2025, the Institute drafted a third version of the S&IP, meant to direct the Institute beyond June 2025 (Years 3-5). S&IPv3, which is a direct response to feedback from our first annual review and site visit as well as the internal reflection and assessment of our leadership team, is also available upon request.

The second quarter of 2025 (April through June) will mark a transitional period for the Institute as we implement adjustments to our research plans and reorganize leadership.

Below we highlight updates between S&IPv2 and S&IPv3.

- **Intellectual Theme.** There are no significant changes to the intellectual theme of ACTION in S&IPv3. The three driving goals of ACTION have not changed.

- **Research Plan.** One change to our research plan in S&IPv3 is the adjusted focus of Thrust SEC-3 toward cyber threat intelligence (CTI). Attribution is still important, but the focus of the thrust now includes sharing intelligence about attacks and vulnerabilities, rather than just attribution.
  The overarching example of the fictional city of New Esperanza has not changed, nor has the structure of GATE or the AI stack.

- **Foundational AI Research.** S&IPv3 includes targeted AI research projects to be tackled over the next program year. We continue to use a structure that defines four AI thrusts and the challenges that define and justify those thrusts. Then we describe research threads (i.e. Proposed Research) that shape our strategy to find solutions to the various challenges. Finally, we present those projects we believe will help further the efforts of the various research thrusts. All projects are collaborative, proposed by a team of ACTION researchers from different campuses. As projects will be the focus of our reporting efforts over the next year, we refer to them as Planned Research.

  We have also removed the various phases from our research thrusts and threads. We considered the phases to be too restrictive and, in some cases, unrealistic. The field of AI is moving too quickly for us to reasonably lay out work phases and timelines beyond the upcoming program year. We believe the way we have identified small, focused projects is a more fluid and productive workflow for conducting fundamental research.

  The main goals and reasoning behind our AI thrusts have not significantly changed. Only the strategy and implementation of those goals has changed. In some cases, we discuss the metrics for determining success of individual projects. When these metrics are not clearly spelled out, there is an implicit understanding that the completion of the specific project is a marker of success within a specific research thread. As discussed in the section on Integration, all projects are vetted and reviewed for their impact to the ACTION's research agenda.

- **Use-Inspired Research.** Our changes to the section on use-inspired research in S&IPv3 mirror those changes to the AI section. Here, we have also removed

phases and timelines beyond Year 3 and have emphasized a structure following (i) Challenges, (ii) Proposed Research, (iii) Planned Research (i.e. projects).

Proposed Research outlines a general vision, and a general strategy for achieving that vision. *Planned Research* refers to the specific projects that will be implemented over the 12-month period of June 2025 - May 2026.

Thrust leads have also revised the general goals of SEC-3 and SEC-4, where obtaining and sharing threat intelligence is expected to play a greater role in the collaborative efforts of the AI agents developed in the AI stack of Thrusts AI-1 through AI-4.

SEC-4 now provides a more-holistic vision for its response and recovery (R&R) and places a greater emphasis on the GATE solution for testing and implementation.

- **Research Risks and Mitigation.** We have expanded the lists of research risks in S&IPv3 to include those risks from our Institute's first annual review that we believe presented the greatest threat. We also included current and planned mitigation, where applicable.

- **Education and Workforce Development.** Changes to Education and Workforce Development in S&IPv3 include the distribution of activities across the Institute.

  We also identify two specific educational activities for Year 3 as well as some subtle adjustments to current and ongoing educational activities. One such example of an adjustment is the inclusion of advanced PhD students into what was our postdoctoral fellowship exchange.

  Where appropriate, we identify timelines for educational activities. We suggest these timelines to help with our own planning for the implementation of the activities, and to ensure we are on track.

  Finally, we have included a list of risks and planned mitigation for our plans for education and workforce development. Many of these risks were either identified during our Year 1 review, or through our experiences with curriculum development over the past year. Awareness of these risks is important to the successful implementation of ACTION's goals.

- **Integration**. We provide additional information about our plans for integration throughout the Institute. We also describe ACTION's new internal request for proposals, and explain our reasoning behind this practice and how we believe it will not only build cohesion and participation among the Institute, but will also allow for greater integration of activities.

- **Broadening Participation.** We streamlined our plans for broadening participation and, at the suggestion of the NSF Site Review Team, bolstered our strategy with support from articles citing evidence-based methods. Primarily, these methods reinforce our efforts to introduce cybersecurity (and thus, Computer Science) at the K-12 level, as well as our efforts to build a community of research enthusiasts that will improve our outreach. We also consider the mitigation of three ongoing risks to successful outreach that reviewers identified in earlier reports.

- **Collaboration and Knowledge Transfer.** We did not make any significant changes to our strategy for Collaboration and Knowledge Transfer.

- **Inclusiveness, Diversity, and Equity Plan.** We modified our strategic plan for inclusivity to ensure we are compliant with current standards for Outreach, providing opportunities for all Americans.

- **Management.** This section continues to reinforce the day-to-day operation of the Institute and the duties of the decision makers and stakeholders. Importantly, this S&IP update names the leadership team of ACTION over the next program year, as well as the reasoning behind the many changes in roles.

- **Intellectual Property.** We did not make any significant changes to our plans for Intellectual Property.

- **Ethics.** We did not make any significant changes to our plans for ethics in AI.

- **Continuity.** We updated the list of key personnel and interim substitutes to reflect the current leadership of the Institute. We also placed a pointer to the section for Continuity in our section for Management.

Despite the revised S&IPv3, our overall research plan remains focused on providing answers to a series of research fundamental, critical questions that refer to the

capabilities required by intelligent security agents. Each research question remains aligned with a dedicated research thrust.

- What is the best way to represent domain knowledge so that learning and reasoning can be performed life-long and at scale with well-defined quality parameters? (Thrust AI-1)
- How can humans interact meaningfully with intelligent security agents to provide required domain knowledge and optimize the assignment of tasks? (Thrust AI-2)
- How can one assess the capabilities of multiple interacting intelligent agents under adversarial influence and in a fast-changing environment? (Thrust AI-3)
- What kind of game-theoretical approaches can effectively model adversarial behavior while providing guarantees on detection effectiveness and the ability to deceive sophisticated attackers? (Thrust AI-4)

Addressing these research questions requires overcoming specific challenges and limitations of current AI approaches and techniques. The results of these efforts will create an AI stack providing a set of capabilities that support novel applications of intelligent agents to cybersecurity.

The novel capabilities provided by the AI stack will inform and revolutionize the way in which the cyber-security life cycle is carried out. Each phase of the now-agent-enabled life cycle becomes the focus of a use-inspired research thrust.

- Thrust SEC-1: The new AI stack will support the creation of a knowledge-driven approach to vulnerability analysis and threat assessment that is target-centric and context-aware.
- Thrust SEC-2: By providing robust learning techniques it will be possible to protect intelligent security agents involved in intrusion detection from tampering and evasion.
- Thrust SEC-3 : The ability to model the behavior and goals of adversaries will provide overall threat intelligence, support the identification of adversarial collaboration, and will also support attack attribution.
- Thrust SEC-4: Inter-agent and agent-human collaboration will support effective approaches to response and recovery where newly acquired knowledge makes the system increasingly resilient to future attacks.

Our revised S&IPv2 and v3 continue to emphasize the importance of our Globally Accessible Testing Environment (GATE) infrastructure, which is a collaborative effort

among the entire ACTION team. Each year, we will encourage efforts to build integrated prototypes, with increasing levels of sophistication, for testing within GATE.

In later years of the ACTION Institute, there will be an increased focus on knowledge transfer, collaboration with industrial partners, and testing of the proposed research ideas in real-world computer systems. We will continue to revisit the S&IP on an annual basis to make sure our strategies are in line with the cutting edge at the intersection of AI and cybersecurity, and in line with the expectations of our sponsors and industry partners.

3.0 Essential Functions of the Institute

3.1 Research

During our second program year, ACTION progressed research along the eight research thrusts as described in the Institute's S&IPv2. With each research thrust being broken in separate threads, and counting GATE as a separate endeavor, there are more than 20 total research/technology activities making up ACTION's research agenda. For the purposes of this report, research "threads," "activities," and "projects" are often used interchangeably. Note that projects integrate with one another by spanning multiple thrusts or through execution in the GATE environment.

Appendix C provides a list of Year 2 ACTION research projects and associated personnel. Appendix D provides details on specific activities over the reporting year for each of the current activities. Most updates are provided in the form of publications or significant progress against stated goals. We take care to highlight contributions of ACTION researchers, both as individuals and as teams, but would like to acknowledge the implied participation of graduate students advised by our co-investigators.

We did not encounter any significant delays or unforeseen challenges to our planned research progress in Year 2. Nor were there any new major research directions. In S&IPv3, however, subtle changes have been made to AI-1, SEC-3, and SEC-4. In that planning document, we place a greater emphasis on reasoning over graphs, rather than the use of Probabilistic Programming Languages (PPL), and on Cyber Threat Intelligence (CTI) instead of attack attribution. These changes came about as products of the research generated by ACTION faculty. We believe these changes provide alternate paths to success rather than significantly impact the direction of our efforts.

The timelines and phases proposed in our S&IPv2 played an important role in launching our research in Year 1. In most cases, the initial phase of each thread spanned 24 months. As such, those phases continue to be reflected in this second annual report.

While reporting on activities during this program year, however, the defined phases did little to advance our research, or to provide a platform for future research efforts. Accordingly, we have removed them from S&IPv3.

3.1.1 Foundational AI research

ACTION's second year of activity can be seen as a continuation of the broad and ambitious projects started in Year 1. Research touched nearly all thrusts and threads, with a greater amount of effort spent on the foundational areas of Thrust AI-1 and SEC-1. As indicated in S&IPv2, this was to be expected.

This program year we also made significant progress on our GATE infrastructure. The leading schools in the GATE efforts have been UC Santa Barbara, Purdue University, and Georgia Tech, with each campus contributing to different aspects of the project. Highlighted achievements include the incorporation of NPCs to generate realistic background traffic [1], and Man-in-the-Middle proxies [2] to add value to traffic logs. We have added a new command line interface for a better user experience and designed a starter kit to help explain the interplay between our GATE and our AI Stack. We also created a demonstration that showcases an attack scenario leveraging CVE-2024-4577 [3] and have begun exploring collaborations with similar technologies, such as USC's SPHERE.

Below, we describe some of our most successful research projects, organized by research thrust. Our metric for success within each thrust is a factor of the publication count and productivity of each thread/project.

Thrust AI-1: Learning and Reasoning with Domain Knowledge

Our most successful projects this year under Thrust AI-1 were those involving advanced knowledge representation learning for logical reasoning. This thread led to ten publications, including those by **Dr. Bo Li** [4-6], **Dr. Dawn Song** [7], **Dr. Stuart Russell** [8-10], and **Dr. Bing Liu** [11-13].

Dr. Bo Li, Dr. Dawn Song, Dr. Stuart Russell, and Dr. Bing Liu all laid the theoretical and practical foundations for AI systems that not only represent knowledge in robust

and interpretable ways but also reason with it effectively and ethically across diverse, evolving tasks.

Most notably, the AI research from the past year has positively impacted development on an early version of ACTION's AI stack, available on GitHub [14], which provides comprehensive tools for agent development and significantly reduces coding overhead for researchers.

The other two research threads in AI-1 led to a number of additional publications. On the subject of knowledge discovery through human-in-the-loop and multi-agent systems, **Dr. Ambuj Singh** produced two papers [15, 16]. Our third thread, veridical secure and private learning with end-to-end quality control, yielded papers by **Dr. Bo Li** and **Dr. Bin Yu** [17-21] and **Dr. Ambuj Singh** [22], and has additional, yet unpublished, contributions from **Dr. Sewoong Oh**.

Graphs also played a larger role than anticipated in building the knowledge base, a shift reflected in S&IPv3's greater emphasis on reasoning over graphs.

Thrust AI-2: Human-Agent Interaction

Our work on designing adaptive and behavior-aware AI to support human decision-making yielded papers by **Dr. Ming Yin** [23-26] and **Dr. Gang Wang** [27].

Developing principled methods for effective human-agent interaction with agent-in-the-loop decision making will impact a number of the security thrusts, especially SEC-4, Response and Recovery.

Thrust AI-2 presents two other significant projects related to human-agent teaming. **Dr. David Wagner** has been experimenting with human+ML systems for detecting jailbreak attacks against LLMs. **Dr. Radha Poovendran** has presented an inference-time defense against backdoor attacks on generation tasks for LLMs [28]. **Dr. Bing Liu** made significant advances on learning through interaction with humans [12, 29-31]. Finally, Dr. Ming Yin and Dr. Ambuj Singh have begun a collaboration to explore how LLMs can serve as facilitators within human-AI groups to improve group decision making outcomes.

Thrust AI-3: Multi-Agent Collaboration

Research on resilient multi-agent reasoning was among the most successful within the AI-3 thrusts. In fact, we have listed **Dr. Jie Gao's** work [32, 33] on how groups of agents discern truth as one of our Institute's Year 2 research highlights. **Dr. Radha Poovendran**, **Dr. Pieter Abbeel**, and **Dr. Ambuj Singh** also use this thread to demonstrate learning (and threats to learning) within networks of agents [34-36].

Dr. Jie Gao and Kevin Lu's work on enabling asymptotic truth learning in a social network [32] will impact Thrust AI-3 and SEC-4. On the surface, the research provides a framework for enhancing collective decision making processes, ensuring that groups can effectively discern the truth even when individual members have limited or imperfect information. We note how the research translates to other networks, including how we can understand reasoning within a multi-agent network of cyber defense agents.

ACTION researchers have also been successful in their research on the modeling of multi-agent networks and robust and adversary-aware reinforcement learning. These threads have led to a number of publications by **Dr. Jie Gao** [32, 37], **Dr. João Hespanha** [38], **Dr. Ninghui Li** [39], **Dr. Jules Chenou and Dr. Tonya Fields** [40], and **Dr. Xiaoming Huo and Dr. Wenke Lee** [41].

Thrust AI-4: Strategic Gaming and Tactical Planning

Dr. João Hespanha's research [42] collected under our thread of adversarial modeling and certification was well-received among ACTION stakeholders. We thereby listed it as one of ACTION's research highlights for Year 2.

Dr. Hespanha and his graduate student, **Sean Anderson**, introduce a method that helps agents continue learning and improving, even in an unpredictable environment. The work will impact AI-3 and SEC-4 in future years.

Dr. Radha Poovendran and Dr. Wenke Lee have collaborations on the integration of knowledge and information flow graphs for enhancing cyber AI games [43, 44]. **Dr. Poovendran** also has ongoing work on the design of defense mechanisms against backdoor attacks on deep neural networks (DNNs) [45, 46]. Thrust AI-4 also includes **Dr. João Hespanha's** additional work on algorithms for modeling deception and other strategic gaming within a security system [47].

Globally Accessible Testing Environment (GATE)

By June 2024, for the first year review, ACTION performers had developed a compelling demonstration of GATE, with a basic virtual private cloud (VPC) design and a prototype of a knowledge base and agent communication channels. The demonstrated system included the automated deployment of the cyber-range infrastructure with Terraform, and we could launch an attack within the system. Prototypes of agents used the preliminary communication stack to exchange information between VPCs to prevent subsequent attacks.

Immediately following the demonstration, we used the GATE infrastructure as a platform for innovation in education. A launched instance of GATE served as a learning testbed to provide visiting summer interns with *hands-on* experience to learn about network traffic, communication protocols, and automated infrastructures. Three undergraduate students and one master's student from around the country came to UCSB and spent the summer learning cybersecurity with GATE. Specifically, these students received insights into real-world security incident response as they gained experience installing and configuring intrusion detection (and intrusion prevention) software [48, 49] that could be tested in realistic attack scenarios.

Being used for educational activities, rather than development, our code commits for GATE were less frequent in July and August 2024. September through December 2024 revealed a period of intense activity, however. We focused on related GATE projects, including the relevant server, the AI-stack [14], and other supporting projects [50]. These parallel initiatives targeted prototyping project improvements, which we will elaborate upon in subsequent sections.

In February and March, we aimed to coalesce the partner institutions towards a unified vision of the compelling demonstration of GATE. We learned two important lessons during this time: the first involved software architectural considerations; and the second involved fostering collaboration.

Architecturally, we discovered that certain development activities will likely trigger alerts on campus threat management systems. This discovery also led to a shift in the architecture of the demo. For example, malicious artifacts cannot be staged on a researcher's machine at institutions with centralized, aggressive threat-management systems – otherwise, the threat management software will proactively delete them before they can be deployed.

Second, we learned about engaging new students in a highly complicated research project. They found the complexity daunting and didn't know where to start, so we decided to make a small, easily deployable project to showcase the agent-based communication aspects of GATE. We envisioned a starter kit to help researchers experiment with the interplay between the GATE server and the AI-stack [14]. The starter kit should have only three primary requirements: Docker, Python, and an OpenAI key.

3.1.2 Use Inspired Research

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

Our thread on knowledge-guided vulnerability detection has been our most successful project in Thrust SEC-1 this year. **Dr. Giovanni Vigna and Dr. Christopher Kruegel** showed a framework designed to enable reproducible exploitation of Common Vulnerabilities and Exposures (CVEs) [51]. **Dr. Bo Li and Dr. Dawn Song** published papers on vulnerability assessment tools using multi-agent methods [5, 52, 53]. **Dr. Z. Berkay Celik** presented novel agents for attack and defense in two papers [54, 55]. **Dr. Radha Poovendran and Dr. Bo Li** investigated vulnerabilities in LLM safety alignments in one paper [56]; and then Dr. Poovendran built upon another of Li's papers [57] in another investigation of inherent LLM API vulnerabilities [58]. And finally, **Dr. Dongyan Xu's** work (currently under development) on improving decompilation tools plays a key role in system security posture assessment.

The projects identified above all contribute to the improvement of autonomous agents' abilities to assess threatening situations and potential vulnerabilities and exploits. This research utilizes the AI thrusts to inform the other SEC thrusts.

Dr. Gang Wang, Dr. David Wagner, and Dr. Dawn Song have all contributed extensive work [59-61] to the construction of a knowledge base for threat assessment. Also, **Dr. Elisa Bertino**, in collaboration with Cisco Systems, published research on continual and system-wide vulnerability assessment [62]. We feature Dr. Bertino's work, titled "Graphene," as one of our research highlights.

Thrust SEC-2: Intelligent Agents for Detection

Our thread for more robust agents for detection has been a highly successful source for mini-projects within Thrust SEC-2. **Dr. Dongyan Xu's** research addresses shortfalls in current approaches to classify diverse malware. He created a framework that automates binary analysis to produce standardized datasets ready for AI/ML modeling. More specifically, his MABEL v2.0 dataset revealed enhanced features for NLP, transformers, and deep learning models.

The generation of new datasets, especially those from real-world research, will improve the detection capabilities of our agents. Dr. Dongyan Xu also submitted the MABEL dataset to NAIRR, providing opportunities for greater impact to the community.

Dr. Sewoong Oh contributed to attribution research through the identification of new attack surfaces on LLM training data, as well as on threat evolution. **Dr. Elisa Bertino** also contributed significant work on solutions to concept drift [63]. **Dr. Z. Berkay Celik** and **Dr. Giovanni Vigna and Dr. Christopher Kruegel** also published novel detection agents for large-scale security incidents [64-66].

Thrust SEC-3: Intelligent Agents for Attribution (and Threat Intelligence)

Dr. Dongyan Xu contributed AI-stack tools to assist humans in performing attribution.

Dr. Dongyan Xu's group has incorporated several Natural Language Processing (NLP) models to enhance malware family attribution and characterization. They believe establishing malware intra-family clusters will enhance overall malware family classification and will use the malware sub-family classification to enrich their meta-model performance. They have also established a workflow for the NLP models to classify malware into malware family variants, classify malware, and characterize malware. This malware characterization routine groups similar malware families together and labels new malware according to API Import Functions, thereby allowing their models to characterize malware based on behavior. Dr. Xu is currently evaluating performance of these NLP models.

In Year 3, via S&IPv3, we have expanded the focus of SEC-3 beyond mere attribution to also include elements of threat intelligence. While attribution is important, the revised SEC-3 will allow our agents to learn and share as much as possible about detected threats and vulnerabilities.

Thrust SEC-4: Intelligent Agents for Response and Recovery

Work in the thread of automated and robust attack recovery has generated publications from **Dr. Z. Berkay Celik** [67], **Dr. Elisa Bertino** [68-70], and **Dr. Wenke Lee** [71].

The work provides insight into the behaviors behind realistic attacks, including systems for predicting and defending against future intrusions, as well as for post-attack forensic analysis.

Dr. Wenke Lee has also contributed work toward response to rare events and new attack plans. His tool for resolving complex semantics has led to the discovery of 22 hidden vulnerabilities – 13 of which were zero-day (including nine issued CVEs, with five rated as critical) [72]. **Dr. Dongyan Xu** and **Dr. Z. Berkay Celik** are currently working on the SecXP framework, a self-explainable neural network architecture designed for security applications.

For Year 3, we have refocused SEC-4 to place a greater emphasis on response against multi-step attacks and sophisticated adversarial manipulations. We believe this research will allow our agents to become more robust in their response and recovery.

3.1.3 Benefits of use-inspired research

Dr. Gang Wang is the faculty advisor for Team PurpCorn-PLAN, one of 10 teams selected to participate in Amazon's Trusted AI challenge [73]. The challenge is focused on the security/safety of coding LLMs. The team is committed to "developing advanced code models that prioritize safety, security, and trustworthiness in modern software development. By integrating interdisciplinary expertise, we strive to create AI systems that generate reliable code while proactively identifying and mitigating potential risks" [74]. **Hadjer Benkraouda**, a member of ACTION's Student Advisory Council also participates in the UIUC Trusted AI challenge team.

Dr. Giovanni Vigna, along with **Dr. Christopher Kruegel**, and **Dr. Wenke Lee** advised teams in the DARPA AI Cyber Challenge (AIxCC) [75]. Their teams, namely the Shellphish Support Syndicate [76] and Team Atlanta [77], make up two of only seven teams selected for the grand challenge. Announced at Black Hat USA in Las Vegas, in August 2023, AIxCC is a two-year competition that asks teams of researchers to develop

AI-powered cyber reasoning systems (CRS) that can automatically find, understand, and patch vulnerabilities in open-source software, without human intervention. Many of UCSB's contributions to the SEC research of ACTION overlap with their contributions to AIxCC. Shellphish and Team Atlanta will return to Las Vegas in August 2025 to showcase their upgraded systems as part of DEF CON 33.

We list the above two team challenges as benefits of use-inspired research because the success found by the challenge participants stems directly from the research underway in the ACTION Institute.

3.2 Nurture and grow the next generation of talent

The ACTION Institute's efforts promoting education and outreach did not deviate from what we described in S&IPv2. At each stage of education (e.g. pre-college, undergraduate, graduate, postdoctoral), our education activities include novel activities, including capture the flag (CTF) experiences and courses focusing on AI for security and security for AI.

3.2.1. Pre-college education

Capture the Flag exercises as a tool for high school education. The High School track of the International Capture the Flag (iCTF) ran for the second time in November 2024. 16 teams faced 28 challenges, submitted 280 flags, and performed 2,930 incorrect attempts. The top teams were:

1. *we lost* - Dos Pueblos High School - $3,000 Prize
2. Touch Grass - Panther Creek High School - $2,000 Prize
3. 95th/Dan Ryan (CTA Red Kube) - University of Chicago Laboratory Schools - $1,000 Prize

For the high school track, all cash prizes were made courtesy of Activision, Inc.

We advertised the High School iCTF extensively through the WiCS network, as well as through contacts created within the CS Teachers Association [78]. We targeted high schools in the urban areas around our partnering campuses (e.g. Chicago, Atlanta, Los

Angeles). Where possible, ACTION faculty and graduate students visited high schools in person to talk about computer security and capture the flag.

Different than in 2023, we released twelve walkthrough videos on the ACTION YouTube page that provided tutorials on how to solve CTF challenges. We advertised these videos, especially to those high school students new to CTF, on a designated Discord server.

Videos for iCTF 2024 remain available on ACTION's YouTube channel. Other training materials are available for download on GitHub.

High School Education beyond iCTF. Starting at the end of September, one UCSB PhD student has been working with teachers and students at Santa Barbara High School, enacting the plans for pre-college education laid out in previous ACTION reports. The student met with students in-person every other week with the goal of broadening computer science and computer security education. The student has kept notes to eventually prepare an experiential paper we hope will provide insight into high school CS education, and will further inform the creation of a curriculum that can be transplanted at other high schools. We have begun planning a teacher-education workshop, scheduled for June 21 at UCSB, where we will share ideas for a computer security curriculum with local teachers.

Following their introduction to computer security research and capture the flag exercises through iCTF, the students at SB High School also participated in Carnegie Mellon's PicoCTF in March 2025.

Classroom training in computer security. The video walkthroughs described above serve as our classroom training for high school students in Year 2. Although we do not yet believe ourselves mature enough to impact the actual in-class training of computer science, we have begun our process of creating on-campus, student-run hacking clubs. We discuss this in more depth below, in the section on Broadening Participation.

3.2.2 University, College, and Community College Education

Capture the Flag exercises as a tool for undergraduate education.

iCTF at the undergraduate level is one of ACTION's primary vehicles for nurturing and growing the next generation of talent. Running from November 16-22, the 2024 iCTF

brought together undergraduate teams from around the world to test their cybersecurity skills. Players competed to solve a variety of challenges, which included hacking websites, fooling AI agents, reverse-engineering models, and fuzzy geographic feature search. ACTION faculty and graduate students developed challenges through collaborative efforts spanning the eleven campuses of the Institute. Every challenge was solved by at least one team. Recruitment and advertising for iCTF was managed through the network of networks managed by the UCSB chapter of Women in CS (WiCS).

In the Undergraduate Division of iCTF, 25 challenges resulted in the submission of 284 correct flags and 2,318 incorrect submissions. 27 teams finished with points. The top teams were:

1. *Hack_of_all_trade*s - University of California, Santa Barbara (UCSB)
2. *Shomiti* - Bangladesh - University of Engineering and Technology
3. *broccoli cheddar soup* - UCSB & Arizona State University (ASU)

The teams from UCSB and ASU were entirely unaffiliated with ACTION PIs. A complete list of teams and full results are available on the undergraduate scoreboard. [79]

We aim to grow the competition in future years through continued advertisement and the release of additional training materials and videos.

We have also considered the creation of a Women in Cybersecurity (WiCyS) program in parallel with WiCS, but we have yet to act on these considerations.

New graduate and undergraduate courses. This program year, we continued our collection of information about course offerings at the various ACTION partner institutions at both the graduate and undergraduate level. ACTION faculty and students have built a list of past, present, and future classes related to these subjects taught across all participating institutions. We are using the resources in place to find potential collaborations and create new material that can be released both online, to the general community, as well as through formal instruction at the various partner institutions. This is an ongoing project, sure to continue into future reporting periods.

Dr. Giovanni Vigna taught a new graduate course on AI+Security at UCSB, and a new undergraduate course that tested instructional modules available through pwn.college [80].

Dr. Dawn Song's new LLM Massive Open Online Course (MOOC) (see below on Workforce Development) was also offered as a course at her university. She is looking into the development of another MOOC in Year 3, with greater connection to ACTION.

Summer research experience for undergraduate and graduate students. Although it is also an outcome of education, we discuss the ACTION Summer Research Experience below in the section on Broadening Participation.

Graduate seminar series. Our Graduate Student Seminar Series, organized by the SAC, provided a forum for invited lecturers to present on cutting edge research. The SAC selected speakers according to the interests of its members. The seminar series also gave senior-level PhD students an opportunity to practice job talks and present on their own material.

Postdoctoral Scholar Training. We promoted a postdoctoral training program, as described in S&IPv2, among ACTION partnering campuses. Like we did with the undergraduate and graduate internships, we set up a secure portal for accepting and managing applications.

The postdoctoral training program has not attracted the attention we expected. An assessment of the program, led this year by the SAC, yielded a realization that many postdoctoral scholars are unable to take six months of time away from their home campus. Their contracts are generally for one-to-two years, which means that, by the time they get situated in their appointment and apply for the position, they're often already looking at their next career move. The SAC proposed that senior PhD students also be allowed to apply for the fellowship/exchange. Senior PhD students (i.e. students who have advanced to candidacy) are often in a longer-term arrangement with a faculty advisor than are PhD students. They also serve to benefit significantly from networking and collaboration.

Moving forward, we have taken the SAC's advice. This adjustment to our educational plans is further described in S&IPv3. Our postdoctoral mentoring plan has not changed otherwise.

3.2.3 Workforce development

Dr. Dawn Song with the Berkeley Center for Responsible Decentralized Intelligence (RDI) [81] launched their LLM Agents MOOC in Fall 2024, focusing on LLM fundamentals, automation capabilities, and agent development infrastructure [82]. Dr. Song developed the MOOC with a Research Scientist at Google DeepMind, and arranged for a dozen guest speakers from leading organizations in the space of AI. The course saw engagement from approximately 14,000 enrolled learners, including 400+ Berkeley students and around 7,000 active Discord participants. The inaugural lecture attracted over 38,000 views, demonstrating the significant demand for advanced AI education.

The course includes these topics:

- Foundation of LLMs
- Reasoning
- Planning, tool use
- LLM agent infrastructure
- Retrieval-augmented generation
- Code generation, data science
- Multimodal agents, robotics
- Evaluation and benchmarking on agent applications
- Privacy, safety, and ethics
- Human-agent interaction, personalization, alignment
- Multi-agent collaboration

Dr. Dawn Song with Berkeley RDI organized a large-scale hackathon that drew more than 2,500 developers. [83] The event secured substantial industry backing, with OpenAI, Google AI, AMD, Intel, and other leading AI organizations contributing to a prize pool exceeding $200,000 in awards, resources, and credits. The hackathon features a judging panel comprising prominent AI experts, reflecting Berkeley RDI's commitment to fostering innovation in AI development.

Dr. Dawn Song's MOOC and hackathon described above relate to ACTION's goals via their subject matter; however, they are not products of ACTION. In Year 3, Dr. Song has expressed an interest in developing a MOOC that is more aligned with security agents. Additional details about this MOOC will be included in future reports.

3.2.4 Informal Learning and Education

Community building through GRIT. In August, **Dr. Christopher Kruegel** presented "AI for Security, Security for AI" as part of 2024's series of Ground-breaking Research/Innovative Technology (GRIT) talks at UCSB.

The GRIT talk series began as part of UCSB's Summer Research Mentorship Program for high school students but quickly grew in popularity. The presentations are now available to the community and can be watched asynchronously on UCTV. [84]

In each GRIT Talk, notable faculty at UCSB share their innovative research in a way that grabs the interests of a general audience as well as experts in their field. Each talk is a concise 28 minutes, followed by a half-hour Q&A where attendees can engage with the presenters directly.

3.3 Broadening participation

The Institute's activities toward broadening participation have been consistent with the plans described in S&IPv2. Specifically, we have focused our efforts on using the tangibility of cybersecurity training (e.g. CTF challenges) and AI (e.g. interfacing with LLMs) to increase accessibility to computer science and cybersecurity education. Our efforts have impacted participation in computer science by members of traditionally underrepresented communities.

Our three primary strategies for broadening participation this year were the iCTF, in-person outreach at local high schools, and our summer internship program. We also attempted multiple explorations into collaborating with Minority Serving Institutions (MSI) on ExpandAI concept designs and proposals.

3.3.1 International Capture the Flag

As noted above, in the Section on Nurturing the Next Generation of Talent, we expanded high school participation in iCTF in 2024. While we did not meet our goal of 20 participating high school teams, our participation count was six higher than in 2023. Note that we had twenty teams register for the CTF, but only 16 ended up playing. The four non-playing teams did not complete post-game surveys, but we believe they

simply had conflicts. As we discuss below, overbooked schedules are a common problem among many high school students.

Help from Women in Computer Science. Although we gave in-person presentations at a number of high schools near ACTION campuses (Santa Barbara, Chicago, Urbana-Champaign, Norfolk, and Atlanta), we owe thanks to the UCSB Chapter of Women in Computer Science (WiCS) for their help with outreach. Through their organizational network that spans the ACTION Institute, they once again took the lead on outreach to high schools and colleges for iCTF participation. The inclusion of WiCS in our outreach has allowed us to connect with various Girls Who Code and Women in Cybersecurity (WiCyS). We will follow up on those new connections at the beginning of Year 3.

3.3.2 Outreach in High Schools

The ACTION Institute has successfully begun in-person outreach to high schools. On September 10, 2024, **Stijn Pletinckx,** a PhD candidate in UCSB's Computer Science Department, delivered a lunch-time presentation to students at Santa Barbara Senior High School, formally launching a collaboration between the SBHS Computer Science Academy and the ACTION AI Institute. Stijn's presentation, titled "Hacking for Good: Unlocking Cybersecurity and Ethical Hacking Careers," was meant to catalyze the creation of a new student-run, hacking club at the public high school. It was the first step in one of ACTION's greater outreach efforts.

Working with the Director of the SBHS CS Academy, Stijn's goal was to use hacking exercises to teach computer science and programming to high school students. With a focus on general problem-solving, his curriculum relied heavily on challenges inspired by Capture the Flag. Stijn helped the SBHS students build a club, meeting with them in person on a bi-weekly basis to provide educational materials. A desired outcome of the collaboration was to reach students who have yet to be fully engaged with the Computer Science Academy.

One aspect of Stijn's plan for assessment was to track performance in organized CTF throughout the academic term. The SBHS team participated in ACTION's iCTF in November 2024 and then in Carnegie Mellon's PicoCTF in March 2025. During the iCTF, which was their first venture into CTF, the SBHS team finished in the top 63%. During PicoCTF, they finished in the top 25%.

A later outcome of this SBHS + ACTION collaboration is a replicable curriculum that can be brought, first to other high schools near ACTION Institute campuses, and subsequently to high schools around the United States. We are actively pursuing the creation of additional clubs in Santa Barbara County. SAC members have also planted seeds for clubs in high schools in Chicago and Urbana, Illinois, and in Norfolk, VA. For example, in March 2025, **Hadjer Benkraouda** discussed the potential for learning cybersecurity through capture the flag challenges as part of UIUC's SAIL, an entirely student-run organization that hosts an annual two-day outreach event for high school students. Student feedback from Hadjer's presentation was positive, and consistent with other evaluations.

Norfolk State will utilize aspects of our curriculum in their Summer 2025 high school summer camps. The University of California, Santa Cruz has borrowed from our curriculum in the development of their future weekend outreach events

When thinking about hacking clubs as future high school outreach from ACTION, this past year has introduced some valuable lessons. First, a motivated and dedicated student leader is essential to the success of the club. High school students seem to be busy and over-scheduled. At least one student in the club needs to be endowed with the responsibility to lead the group; that person has to be excited about the material. Second, regarding students' schedules, a fixed meeting time is very important, especially during the club's first few weeks. We held our club meetings over the lunch period, which was the norm at SBHS for student-led clubs. A 30-minute period for instruction was limiting, but finding another time that worked for all students was also challenging.

3.3.3 Summer Research Experiences

Internships for high school students. **Dr. Christopher Kruegel** hosted a high school student intern from June through August 2024. She collaborated with a graduate student at UCSB on a project related to AI and computer security.

Dr. Christopher Kruegel will host another high school student from Northern California over summer 2025.

iCTF 2024 led to a connection between the Laboratory High School near the University of Chicago. **Dr. Nick Feamster** is now working with the Computer Science teacher to develop a CTF team for future years, and also to build a pipeline for internships within his lab.

Internships for undergraduate and graduate students.

Summer Research Experiences at UCSB.
2024 marked the inaugural year of our summer program. ACTION selected a group of 15 students for a research experience at UC Santa Barbara. There were eight colleges and universities represented:

1. California State University, San Bernardino
2. De Anza Community College, California
3. Cornell University, New York
4. University of Minnesota
5. Norfolk State University, Virginia
6. American River College, California
7. Bakersfield College, California (but transferring to UC Berkeley)
8. University of California, Santa Barbara

Participants formed four separate research teams. Each project team consisted of one graduate student, two undergraduates, and a mentor who was either a senior-level PhD student, a postdoctoral scholar, or a software engineer. Each of the projects contributed to ACTION research. More specifically, one project helped with development in GATE; the others dove into fundamental research.

Summer participants presented their four group projects as collaborative posters at UCSB's 2024 Undergraduate Research Symposium. In total, there were 65 posters on display at the Symposium, spanning engineering, the sciences, and the social sciences. The event marked the end of an eight-week research experience, full of hands-on training, professional development, and networking opportunities for our scholars. ACTION was one of 15 educational programs participating in the activity.

Summer Research Experiences Beyond UCSB. In addition to the undergraduate projects supported by ACTION at UC Santa Barbara, Norfolk State University and Rutgers University held their own summer internship programs.

NSU's 2024 Summer Research Program, run by the College of Science, Engineering, and Technology, saw more than 40 young scientists participate in courses on ethics in research, technical writing, and professional development. Specific to ACTION, **Dr. Tonya Fields** advised two undergraduate students, **Wendon Doswell and Clay Jones,** in a project incorporating NLP techniques for creating LLMs.

Dr. Fields and Dr. Hoppa co-advised another student, **Korey Bailey**, on a project considering the implications of ChatGPT for Cybersecurity. They presented the resulting paper to the 2024 International Conference on Computational Science and Computational Intelligence (CSCI'24).

Dr. Jules Chenou, a research scientist in the NSU's Cybersecurity Complex, mentored **Jasmine McDuffie** in an undergraduate project titled "Self-Adaptive Petri Net for Detection of Malware Intrusion." Jasmine provided a technical report and made an enthusiastic promise to build on her efforts with future research.

Dr. Chenou also advised an NSU graduate student, **Timothy Agboada**, in a project that studies an intrusion detection system for identifying and mitigating potential cyber attacks against a cyber-physical system. In this project, Agboada and Dr. Chenou use UPPAAL to model the autonomous braking systems of vehicles, with hopes of providing a framework for enhancing and improving the safety of the systems.

At Rutgers University, four REU students worked on topics related to ACTION in Summer 2024. Two of the students, **Filip Uradník and Julia Krizanova**, hailed from Charles University and were funded by the Czech Republic. The other two students, **Amanda Wang and Rhett Olson** were from Princeton University and the University of Minnesota, respectively.

At UCSB, unrelated to the summer research experience, another undergraduate collaborated on a program related to AI and Cybersecurity in VR. The project came about through an outreach presentation at a meeting hosted by the UCSB Undergraduate Research and Creative Activities and the local chapter of the Society for the Advancement of Chicanos/Hispanics and Native Americans in Science (SACNAS). The project could help inform some aspects of agent learning (Thrust AI-1), human-agent collaboration (Thrust AI-2), and assessment (Thrust SEC-1).

3.4 Multidisciplinary integration and community building

3.4.1 Seminars

Like in our first year of operation, ACTION continued the practice of community building through the planning and implementation of a series of cross-campus seminars. We offered these lectures as hybrid events and advertised them across all

campuses in the Institute, as well as to sponsoring organizations. ACTION hosts two types of seminars: (i) distinguished lectures for experts in the field; and (ii) student seminars for new faculty and senior graduate students. The SAC makes decisions about whom to invite, helps create the schedules, and generally hosts the talks. ACTION management hosts recordings of the distinguished lectures on the ACTION YouTube Channel.

Below is a list of 2024-2025 seminars.

Distinguished Lectures.

- Dec 5, 2024. "The Challenges of Machine Learning in Adversarial settings: A systems perspective ." Patrick McDaniel, University of Wisconsin-Madison.
- Feb 10, 2025. "Enabling and Evaluating Human-Agent Collaboration." Diyi Yang, Stanford University.

Student seminars.

- Sept 26, 2024. "Data-driven analytics to empower security and protection in the AI era." Shuang Hao, University of Texas, Dallas.
- Jan 30, 2025. "Weaving an intelligent net with more connected, accessible, and reliable machine intelligence." Shinan Liu, University of Chicago.
- May 9, 2025. "Data-driven security operations - The advantages and limitations of AI." Thijs van Ede, University of Twente, The Netherlands.

The SAC also uses the student seminars to discuss professional development activities and practice job talks.

3.4.2 Campus Visits for Community Building

To help with community building among graduate students and faculty, the SAC developed an integration plan during their February meeting. In summary, small travel grants would be made to SAC members (one per member per year) to visit other Institute campuses to deliver in-person talks and become familiar with current research throughout the Institute. SAC members need only propose the trip for it to be approved. We expect these collaborative meetings to occur in the next program year.

3.4.3 Other Meetings and Communication

ACTION maintains a website and a professional Slack workspace for sharing information and announcements. We also hold monthly Institute-wide meetings for faculty.

Dr. Giovanni Vigna and Dr. David Evans, along with the SAC, have organized a weekly research group meeting where participants share brief overviews of various papers connecting to the area of knowledge behind each of the use-inspired thrusts. The meetings are popular, being loosely structured to give students free-reign to collect papers. With participants hailing from all ACTION campuses, these meetings are an example of the community building efforts of the Institute.

We ask that thrust leads also organize smaller meetings among their respective performers on a monthly or quarterly basis. ACTION management offers assistance with scheduling, should thrust leaders need it.

3.5 Multi-organizational synergies/achievements

In August 2024, **Dr. Xiaoming Huo** (Georgia Tech) and **Dr. Bin Yu** (UC Berkeley) organized an invited session at the Joint Statistical Meeting (JSM 2024), on the topics of "statistical methods in cyber security." Participating ACTION faculty included **Dr. Jie Gao** (Rutgers), **Dr. Bo Li** (U Chicago/UIUC), **Dr. Radha Poovendran** (U Washington) and **Dr. João Hespanha** (UCSB).

Dr. Radha Poovendran continued to lead strong collaborations with other ACTION institute partners, including Dr. Wenke Lee from Georgia Tech and Dr. Bo Li from U Chicago/UIUC. Other collaborations included Dr. B. Y. Lin at AI2, Prof. Y. Choi at Stanford University, Prof. L. Niu at University of Washington, Prof. S. Moothedath at Iowa State, and Prof. B. Ramasubramanian at Western Washington University.

Dr. Giovanni Vigna received the SIGSAC Outstanding Innovation Award at ACM CCS in Salt Lake City in October 2024. "This award is given for outstanding and innovative technical contributions to the field of computer and communication security that have had a lasting impact in furthering or understanding the theory or development of secure systems." Previous ACTION research faculty to win this award have been Dr. Dawn Song (2020), Dr. Wenke Lee (2019), and Dr. David Wagner (2016).

The 2024 iCTF involved collaboration among the ACTION Institutes. During the month preceding the security exercise, the six participating ACTION institutions familiar with security exercises generated 28 high school challenges (+17 from 2023) and 25 undergraduate challenges (+11 from 2023). Each institution worked with a lead graduate student from UCSB to create and deploy their challenge. The iCTF was played by 27 undergraduate and 16 high school teams from across the country.

One result of the research group meetings led by **Dr. Giovanni Vigna**, **Dr. David Evans**, and the **SAC** described above in section 3.4.3 has been the drafting and submission of a systematization of knowledge (SoK) on AI for threat intelligence. This paper has helped ACTION refocus SEC-3 on Threat Intelligence. We discuss this SEC-3 thrust adjustment in S&IPv3.

We believe the revision of our S&IP to include an agreed-upon RFP for funding to be a significant multi-organizational achievement, a codified demonstration of teamwork among the 11 campuses.

3.6 Knowledge transfer

3.6.1 Publications

Visit https://action.ucsb.edu/knowledge-transfer or the ACTION Google Scholar profile (https://scholar.google.com/citations?hl=en&user=KMn7qeAAAAAJ) for a list of publications.

3.6.2 Presentations

This program year, ACTION faculty presented work described above, in Section 3, and also in Appendices B-D, at various conferences and professional meetings. We provide a few in the list below:

- **Dongyan Xu,** LeHack, Paris, France. July 2024.
- **Xiaoming Huo and Wenke Lee**, Data Science Seminar, School of Mathematics and Statistics, University of Sydney, Australia. July 2024.
- **Xiaoming Huo and Wenke Lee**, Machine Learning @ Georgia Tech Student Conference. August 2024.

- **Giovanni Vigna,** Strategically Leveraging AI in Education Conference. August 2024. (also present was Chad Lane, INVITE AI Institute).
- **Dongyan Xu**, DEF CON, Las Vegas, NV. August 2024.
- **Tim Robinson,** Summit for Digital Trust, Glion, Switzerland 2024.
- **Dongyan Xu,** Sec-T, Stockholm, Sweden. September 2024.
- **Stijn Pletinckx**, "Hacking for Good: Unlocking Cybersecurity and Ethical Hacking Careers," Santa Barbara High School. September 2024.
- **Giovanni Vigna,** Annual SaTC PI Meeting, Pittsburgh, PA. September 2024.
- **Dongyan Xu,** BlackHat SecTor, Toronto, Canada. October 2024.
- **Dongyan Xu,** GrrCon, Grand Rapids, Michigan. October 2024.
- **Dongyan Xu,** HackerHalted, Atlanta, Georgia. October 2024.
- **Xiaoming Huo and Wenke Lee**, SIAM Seminar, School of Mathematics, Georgia Tech. November 2024.
- **Xiaoming Huo and Wenke Lee**, Seminar at the Department of Mathematics, University of Bayreuth, Germany. November 2024.
- **Xiaoming Huo and Wenke Lee**, Sydney Workshop on Mathematics of Data Science, Australia. December 2024.
- **Xiaoming Huo and Wenke Lee**, Joint Meeting of the New Zealand, Australian, and American Mathematical Societies, Auckland, New Zealand. December 2024.
- **Xiaoming Huo and Wenke Lee**, Machine Learning and Optimization Seminar, University of Hong Kong, Hong Kong. December 2024.
- **Bin Yu,** Data Science Seminar, School of Mathematics and Statistics, University of Sydney, Australia. December 2024.
- **Sean Anderson and João Hespanha**, DHS Quarterly Exchange with Swedish Civil Contingencies Agency (MSB), Berkeley, CA. March 2025.
- **Hadjer Benkraouda**, "Cybersecurity Foundations: Learning Through Capture The Flag Challenges," Presentation at SAIL 2025, Urbana, IL. March 2025.
- **Hadjer Benkraouda**, Women in CyberSecurity Annual Conference (WiCys), Dallas, TX. April 2025. (Best Graduate Poster Award)

3.6.3 Other Meetings Related to Knowledge Sharing

ACTION remains involved with the AI Institutes Virtual Organization (AIVO) thereby maintaining connections to the other NSF-funded AI Institutes. ACTION was represented at the Summit for AI Institute Leadership (SAIL) in October 2024 in Pittsburgh, PA.

In November 2024, **Dr. Dongyan Xu** and Tim Robinson met with researchers at the Center for Security and Emerging Technology (CSET), an organization with interest in policy development out of Georgetown University [85]. Dr. Xu presented ACTION research and goals, with an emphasis on the GATE architecture, as well as on education and workforce development.

Also in November 2024, **Dr. Giovanni Vigna** delivered a presentation on ACTION to the Academic and Student Affairs Committee at the meeting of the University of California Regents.

In December 2024, the Networking and Information Technology Research and Development (NITRD) invited ACTION performers to join a panel on Cybersecurity and AI R&D at ACSAC 2024 [86] in Honolulu, HI. This panel discussed the USG research priorities in cybersecurity and AI R&D and is connected with the conference's theme of Advancing AI and Cybersecurity. **Dr. Dongyan Xu** represented ACTION.

In January 2025, **Dr. Dongyan Xu** represented ACTION at a workshop panel on national critical infrastructure cybersecurity risks and opportunities in Miami, FL. [87] Dr. Xu shared his experiences and observations in creating multi-aspect, multi-fidelity digital twins for critical infrastructure modeling, simulation, and security/resiliency assessment, based on his current research efforts under ACTION and other relevant projects. In particular, he discussed with other panelists and the audience the use of AI for the generation and training of digital twin models for performing key functions in the cyber defense life cycle. The connection for participation in this panel was made by DHS.

In March 2025, **Dr. Wenke Lee** joined a panel at the 2nd Annual Billington State and Local Cybersecurity Summit in Washington, D.C. [88] The panel was titled "AI on the Cybersecurity Frontier: Cutting Edge R&D at Academia." Dr. Lee gave an overview of the participating institutions, mission and goals, project selection criteria, connections to state/local governments, impacts, etc. He also discussed how one cannot have blind trust in any single AI agent. Instead, for cybersecurity, we need a multi-modal and multi-agent approach where agents collaborate amongst themselves and with humans to achieve better accuracy and robustness, in particular, in the face of attacks/manipulation by malicious AI agents/adversaries. He stressed how humans need to be the ultimate decision makers.

Cheng Xin and Kevin Lu, a postdoc and student advised by **Dr. Jie Gao**, delivered flash talks inspired by their ACTION-related research at the DIMACS Workshop on Spreading on Social Networks – Theory and Applications, held at Rutgers University from October 21, 2024 - October 23, 2024. The event was co-sponsored by the Center for BioComplexity, High Meadows Environmental Institute, Princeton University and The Santa Fe Institute.

3.6.4 Other Knowledge Sharing

Dr. Radha Poovendran's group made the MAGPIE code repository publicly available on GitHub. MAGPIE enables generation of high-quality instruction datasets at scale for LLMs via self-synthesis. The repo has over 500 stars as of March 2025. [89]

Dr. Dongyan Xu and Solomon Sonya made ACTION's first malware analysis dataset for AI/ML Modeling (MABEL: Malware Analysis Benchmark for AI & ML) available on GitHub [90]. In February 2025, they released MABEL to NAIRR.

ACTION researchers at UCSB, Purdue, and Georgia Tech also made code for the AI-stack and various aspects of GATE available on GitHub [14, 50]. These updates are further detailed in Appendix D.

There are no formal exchanges of IP or technology transfer during our second year. However, as noted in the use-inspired section above, ACTION faculty researchers like **Dr. Elisa Bertino and Dr. Gang Wang** continue collaborations with industry giants like IBM, Cisco, and eBay, as well as NCSA. These collaborations will seed knowledge transfer in later years of ACTION.

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

3.7.1 Collaborative grants

Dr. David Wagner and **Dr. Christopher Kruegel** received a grant AI for Cybersecurity from the UC Noyce Initiative. While the funding program may be internal to the University of California, the grant marks the impact the ACTION faculty have when conducting collaborative research. A third Co-PI on this project is Hao Chen from the University of California, Davis.

3.7.2 Other collaborative efforts

GATE Integration with USC SPHERE. In an effort to expand the use of GATE while also eliminating reliance on a single vendor (AWS), the GATE development team began a partnership with the USC SPHERE [91], whose comprehensive technical documentation has enabled us to build a functional prototype. The current GATE-enabled virtual private cloud (VPC) now successfully demonstrates isolated virtualization platforms exchanging information through a deployed ACTION server with simple agents. As development is ongoing, SPHERE support staff continue to provide expert guidance to ACTION engineers on network architecture challenges.

Dr. Luis Garcia, from the University of Utah, has become a steady collaborator with our ACTION software team. We have worked to integrate Dr. Garcia's high-fidelity PLC traffic simulations into GATE, thereby enhancing our real-world applicability. The ensuing partnership has set up a strong foundation for GATE's deployment, adoption, and utilization in complex virtualized environments. We will continue reporting on updates in future reports.

Community building through Center for Digital Trust. The Center for Digital Trust (C4DT) [92] at the Swiss Federal Institute of Technology Lausanne (EPFL) invited the ACTION Institute to the Second Annual Summit of Centers for Digital Trust in Glion, Switzerland in August 2024. EPFL brought sixteen research centers across five continents to the summit to link the community and share operational best practices. Participants included CMU Africa, KTH Royal Institute of Technology, Sweden, Indian Institute of Technology, Bombay, Uppsala University, Sweden, Korea Advanced Institute of Science and Technology. US organizations included Carnegie Mellon University, Arizona State University, and Purdue University.

Participants in the summit discussed their centers' missions and impacts, and explored new collaborations among future goals. They engaged in working groups on CTF for Education, university cybersecurity, and e-IDs. The summit culminated in a dialogue with international journalists and their experiences with digital trust.

As part of his presentations on ACTION, Managing Director Robinson also spoke about the National Science Foundation's community of 25+ AI institutes. Participation in the summit led to ACTION's connection to researchers in digital security and CPS in Sweden and Scotland, and to high school and undergraduate CTF participants in Kenya

and other African nations. Domestically, the summit led to the inclusion of a new EAB member, and to stronger ties with CMU and ASU, especially with respect to CTF and pwn.college.

The Third Annual Summit for Digital Trust will take place at EPFL in August 2025. ACTION will again participate in this international, community building event.

Foreign collaborations. **Dr. Radha Poovendran** met with faculty from the University of Padua CSE (Prof. Mauro Conti of SPRITZ security research group) in September 2024 to explore AI-security collaborations. He later met with faculty from the KTH, Sweden, security group. **Dr. Giovanni Vigna** met with faculty at the University of Trieste in Italy to also explore collaboration. In May 2025, we will host Dr. Thijs van Ede to discuss collaboration in Year 3 with the University of Twente in the Netherlands.

4.0 Evaluation and Assessment

Again in Year 2, the external evaluation team led by Dr. Tarek Azzam helped with applications for iCTF and the summer research experiences, ensuring that demographic information was properly collected. Evaluators helped with rubrics for student evaluation to help us meet our goals for improving participation in STEM. They also collected survey data and prepared evaluative reports ACTION can use for improvements.

The evaluation team has begun review of the Institute's plan and rubric for graduate student development. Finally, in response to comments from the Year 1 SVT, they have committed to helping the Institute formalize a model for overall Institute assessment.

Dr. Azzam's team continues to meet quarterly with Managing Director Robinson and Director Vigna about program evaluation. Below, in section 4.1, we included an excerpt of their suggestions provided in their own annual report.

4.1. Education and Outreach Evaluation

Overall Institute Assessment.
Overall, the Institute shows promise and growth across key domains in terms of the education and workforce development outreach and mentorship, academic preparation, research preparation, and professional development and networking. The Institute

implemented impactful activities that engaged high school students, undergraduate students, and graduate students. Moreover, many students have had the opportunity to engage in activities that contribute to preparation for academic and professional trajectories within AI and cybersecurity fields.

Moreover, the ACTION Institute has been very responsive to the evaluation plans and is enthusiastic about incorporating more evaluation services into their activities. ACTION management even translated the Institute's Strategic Implementation Plan into a Logic Model. The language and deliverable dates presented in this model are directly related to the SIP. Since the SIP has been rewritten since the initial draft of the associated logic model, ACTION management will work with the evaluation team to revise and refine the SIP logic model to accurately reflect the priorities of the Institute.

Overall, the ACTION Institute shows signs of healthy implementation over the first two years and expressed enthusiasm to further leverage the evaluation team. In particular, the evaluation will expand to address academic preparation, Summer research experiences across multiple institutions, postdoctoral experiences, and strategies to broaden the impact of the Institute. The Institute shows promise in further refining existing activities and using feedback from previous participants to replicate effective practices across affiliated ACTION institutions. Future years of the ACTION Institute are likely to include more evaluation involvement and utilization of evidence for replicating or scaling ACTION activities.

Conclusions.

The ACTION Institute successfully implemented multiple activities that support both the broader mission of the Institute and educational and workforce development components specifically. While the evaluation does highlight the success of the Institute, the Institute also has strong potential for growth in terms of the service and engagement that they provide. This is common for programs that are starting, and the evaluation team strives to support the development of the program by providing additional monitoring to ensure that specific components such as academic preparation are more thoroughly addressed.

The Institute may first consider revisiting the educational and workforce development logic model to assess which activities are being offered and their direct connection to outcomes. Refining the logic model to reflect which activities have been implemented and are currently functioning is important in order to address the activities that have not yet been implemented. For example, the MOOC has not had as much clarity in terms of implementation compared to the Summer research experience. Again, this is

common for programs early on in their development, and the evaluation team plans to support the implementation of activities that are in development by providing useful and relevant feedback.

Second, the Institute plans to revisit the Strategic Implementation Plan logic model to reflect the most current structure of the ACTION Institute.

Finally, the Institute may also wish to increase outreach and participation in broad-access activities such as the iCTF competition. This could also include evaluating the needs of current undergraduate students in terms of academic preparation and identifying populations that would most benefit from services offered by the ACTION Institute. While the Institute has many ideas about what sorts of services to offer, connecting those services with the correct audience will remain a goal of the evaluation team to help the Institute achieve a goal of broadening participation and training the next generation of talent in cybersecurity and AI-related fields. Altogether, the ACTION Institute shows appropriate developmental progress and receives feedback about their implementation well. The positive attitude towards feedback and demonstrated effectiveness from completed surveys and focus groups show signs of progress towards achieving broader impact goals in terms of education and workforce development.

4.2. Response to External Evaluation

We thank Dr. Azzam and his team for their assistance with program assessment. Their help with justifying our outreach efforts through evidence-based research was invaluable, and their comparison of data collected from iCTF participants over years 1 and 2 will allow us to improve our program.

Also, now that we have had time to generate deliverables and short-term outcomes from our education and outreach activities, we believe the logic model developed by the evaluation team could benefit from a revision, and a better division and realignment with the headings of "Nurture and Grow the Next Generation of Talent" and "Broadening Participation." We look forward to working with the team on this assessment tool.

Our S&IP revision required a considerable amount of our time during the past year. With that plan now solidified, we have a better opportunity to work with the evaluators to tune our overall assessment model. For instance, a number of the stated deliverables in our initial draft are associated with published papers. Our plan for the upcoming

reporting periods will be to demonstrate where we have met our proposed milestones and, furthermore, to cite the impact those deliverables have made on the community through reciprocal citations.

5.0 Infrastructure and Management

5.1. Configuration and Leadership

5.1.1 Organizational Chart

During Year 2, the ACTION Institute maintained the organizational structure described in the S&IPv2 and S&IPv3, which is visualized by the organizational chart in Figure 1.

Brief organizational updates follow:

Executive Committee. The EC met weekly for updates on activities and planning. In Year 2 of ACTION, the EC reevaluated committee membership and the various leads for the five organizational pillars (Foundational Research, Use-Inspired Research, Broadening Participation and Outreach, Education, and Industry Liaisons).

Student Advisory Council. The SAC has maintained a monthly meeting time. **Shinan Liu** from University of Chicago served a second term as council head.

SAC members have been very active in Institute activities. They helped with the initial round of summer internship application review in February 2025. They helped select speakers for the quarterly ACTION seminars, inviting them when appropriate. **Council Head Liu** often initiated the hosting of seminars and ran council meetings. The SAC helped with promotion of iCTF 2025. Those members with CTF experience participated in challenge design. The SAC has also been heavily involved in the SoK meetings.

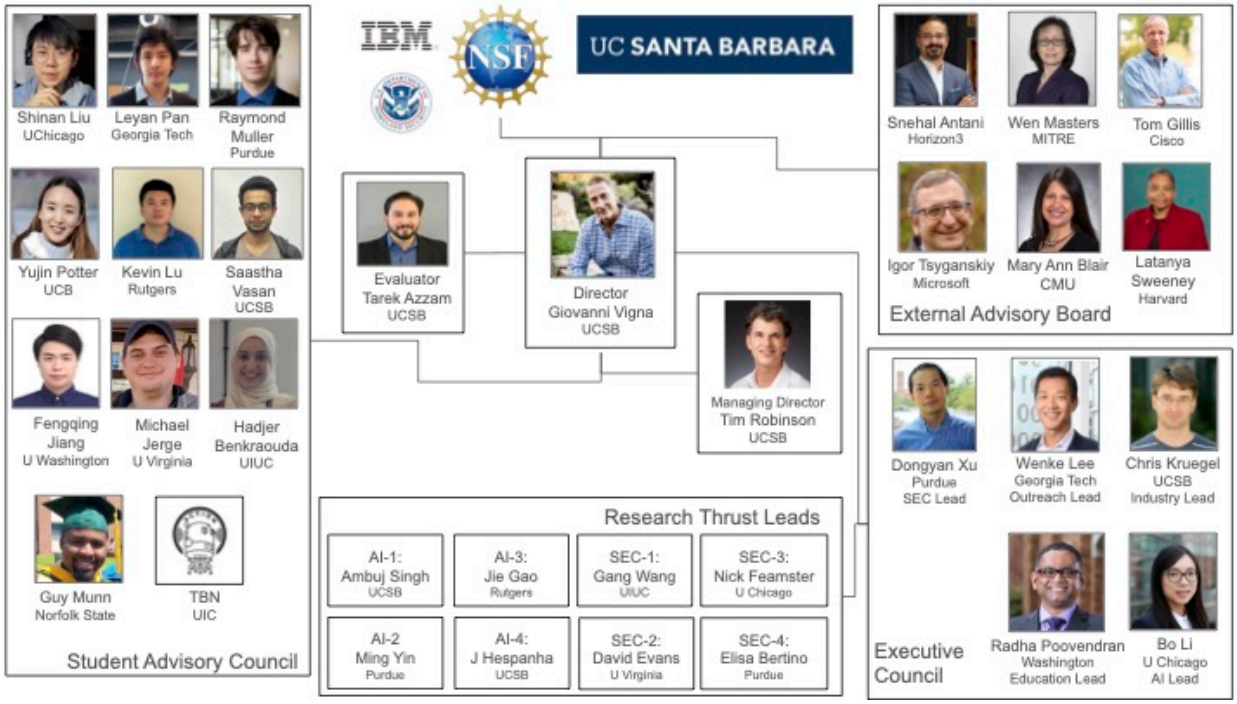


Fig 1. ACTION Institute Management Organizational Chart

External Advisory Board. The EAB maintained their organizational charter (available upon request). They've met twice this year. During meetings, ACTION leadership provides Institute updates and then receives commendations and recommendations from the EAB. All Board members are invited to the ACTION annual review. All Board members receive access to quarterly and annual reports.

At the request of the SVT, we sought the inclusion of an academic CISO to our EAB. Mary Ann Blair joined in late November 2024. She has been an active participant since then.

5.1.2 Ethics Training

All graduate students participating in the ACTION Institute have completed the "Responsible Conduct of Research" (RCR) Training through CITIprogram or a

comparable training. We have yet to explore the inclusion of additional education and training in ethical behavior.

5.2 Management Systems and Reporting

5.2.1 Management Systems

5.2.1.1 Role of advisory boards.

The roles of the EAB and SAC, as described in the S&IP, have not changed during Year 2. Both advisory boards have charters to define their operations. The charters are available to board members upon request. The SAC has been more active than the EAB, having met on a monthly basis.

5.2.1.2 Project management

How is it determined which projects are needed to achieve the Institute's strategic plan?

ACTION's S&IPv2 laid out what faculty researchers believed to be the necessary activities for the first two years of the Institute. These were the foundational activities for research, education, and outreach. This program year, faculty researchers continued these efforts, largely on their own, with support from the relevant thrust leads.

S&IPv3 outlines the plans for future project selection to achieve the Institute's strategic plan. In summary, following the Institute's annual review, the EC and thrust leads will meet with the EAB to assess current institute vision and plans for the subsequent year. The EC will then announce a request for mini-proposals to the ACTION community. Institute faculty prepare proposals for activities based on this RFP. The EC and thrust leads then select those proposals that will have the greatest impact and advance the strategic plan of the institute. All proposals are required to be collaborative, with two or more faculty from different campuses, and must have defined milestones and deliverables. Accepted proposals are folded into the next iteration of the S&IP.

How does the Institute assess the quality and impacts of the projects?

In Year 2, project quality and impact were determined based on publishability and alignment with the S&IP.

Moving forward, project success will be measured against the milestones outlined in the submitted mini-proposals. Assessment and evaluation will occur annually by the EC and, when appropriate, by the EAB, NSF, and other ACTION stakeholders. While our assessment team is primarily responsible for measuring progress along the lines of Education and Broadening Participation, they have also offered assistance in the future assessment of the Institute as a whole.

What role does the advisory board play in project review and assessment?

Currently, the EAB provides commendations and recommendations on high level aspects of the ACTION Institute. As the EAB becomes better-grounded in the Institute, they may begin to have more input on project assessment.

How does ACTION identify associated projects awarded to Institute faculty members' departments that are needed by the Institute to achieve the strategic plan?

We ask ACTION faculty to provide updates on associated projects that contribute to the Institute's strategic plan. The responsibility is that of the associated project PI, but with all ACTION faculty now engaged in an internal RFP, there is an incentive for providing updates on quarterly progress reports.

How was ACTION's research team formed, including the role of faculty/staff from partner and collaboration institutions in research and education?

The ACTION community was initially determined during the drafting of the initial proposal to NSF. Since then, some additional faculty have entered the community.

Initial thrust leads and EC members were also named during the NSF proposal phase. Many of these roles have changed over the current reporting period. Dr. Bo Li, with her expertise in AI, now leads the AI thrusts on the EC; Dr. Radha Poovendran, who is currently engaged in a wealth of educational activities at his home institution, acts as Education on the EC. Dr. Ambuj Singh, an expert on graph mining and human-agent teaming, leads Thrust AI-1. Dr. David Evans and Dr. Nick Feamster lead Thrusts SEC-2 and SEC-3 respectively. They replaced Dr. Christopher Kruegel and Dr. Giovanni Vigna, who already hold leadership positions on the EC. Finally, Dr. Elisa Bertino, who provided the greatest amount of critical input on the refocusing of ACTION's goals for Response and Recovery activities, volunteered to lead SEC-4. She does this in place of Dr. Wenke Lee, who continues with his role as outreach lead on the EC.

In this way, the ACTION management team has become well-distributed across the institute's campuses, and has been initialized by a desire to participate and lead.

Teams within research thrusts are currently left to the interests and collaborations of the individual faculty researchers. We encourage organic collaborations and try to create opportunities for researchers to come together.

> *How does ACTION integrate any other programs or resources into the research program?*

We do not foresee any issues with integrating complementary research. Decisions on how to handle integrations are left to the individual faculty. They are encouraged to seek help from the ACTION EC.

5.2.3 Conflicts of Interest

We do not have any current conflicts of interest. Care was taken to include COI procedures in the S&IP.

References

[1] Carnegie Mellon University Software Engineering Institute, "GHOSTS." [Online]. Available: https://cmu-sei.github.io/GHOSTS/

[2] mitmproxy Project, "mitmproxy." [Online]. Available: https://mitmproxy.org/

[3] NIST, "CVE-2024-4577 Detail," *National Vulnerability Database*. [Online]. Available: https://nvd.nist.gov/vuln/detail/cve-2024-4577

[4] M. Kang et al., "C-RAG: Certified Generation Risks for Retrieval-Augmented Language Models," in *Proc. 41st Int. Conf. Mach. Learn. (ICML)*, Jul. 2024.

[5] C. Guo et al., "RedCode: Risky Code Execution and Generation Benchmark for Code Agents," in *Proc. 38th Conf. Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2024. [Online]. Available: https://redcode-agent.github.io/

[6] Y. Li et al., "Consistency Purification: Effective and Efficient Diffusion Purification Towards Certified Robustness," in *Proc. 38th Conf. Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2024.

[7] E. Pasewark et al., "Re-Tuning: Overcoming the Compositionality Limits of Large Language Models with Recursive Tuning," in *Proc. 62nd Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2024.

[8] L. Lang et al., "When Your AIs Deceive You: Challenges of Partial Observability in Reinforcement Learning from Human Feedback," *arXiv preprint arXiv:2402.17747*, 2024.

[9] A. Garber, R. Subramani, L. Luu, M. Bedaywi, S. Russell, and S. Emmons, "Will an AI with Private Information Allow Itself to Be Switched Off?," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, AI Alignment Track, 2025.

[10] S. Kapur, E. Jenner, and S. Russell, "Diffusion on Syntax Trees for Program Synthesis," in *Proc. 13th Int. Conf. Learn. Represent. (ICLR)*, 2025.

[11] C. Niu, G. Pang, L. Chen, and B. Liu, "Replay-and-Forget-Free Graph Class-Incremental Learning: A Task Profiling and Prompting Approach," in *Proc. 38th Conf. Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2024.

[12] S. Momeni, S. Mazumder, and B. Liu, "Continual Learning Using a Kernel-Based Method Over Foundation Models," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2025.

[13] C. Xiao and B. Liu, "Generalizing Reasoning Problems to Longer Lengths," in *Proc. 13th Int. Conf. Learn. Represent. (ICLR)*, 2025. [Online]. Available: https://openreview.net/pdf?id=zpENPcQSj1

[14] ACTION AI Institute, "AI Stack," *GitHub*. [Online]. Available: https://github.com/action-ai-institute/ai-stack

[15] D. Wang et al., "Global Human-Guided Counterfactual Explanations for Molecular Properties via Reinforcement Learning," in *Proc. 30th ACM SIGKDD Conf. Knowl. Discov. Data Mining (KDD)*, 2024, pp. 2991–3000.

[16] Q. Wu, Z. Chen, W. Corcoran, M. Sra, and A. K. Singh, "GraphEval36K: Benchmarking Coding and Reasoning Capabilities of Large Language Models on Graph Datasets," in *Findings Assoc. Comput. Linguistics: NAACL 2025*, Apr. 2025, pp. 8095–8117.

[17] A. Zhou et al., "KnowGraph: Knowledge-Enabled Anomaly Detection via Logical Reasoning on Graph Data," in *Proc. 2024 ACM SIGSAC Conf. Comput. Commun. Secur. (CCS)*, 2024.

[18] A. Zhou, B. Li, and H. Wang, "Robust Prompt Optimization for Defending Language Models Against Jailbreaking Attacks," in *Proc. 38th Conf. Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2024.

[19] J. Wang et al., "BackdoorAlign: Mitigating Fine-Tuning Based Jailbreak Attack with Backdoor Enhanced Safety Alignment," in *Proc. 38th Conf. Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2024.

[20] A. R. Hsu et al., "Rate, Explain and Cite (REC): Enhanced Explanation and Attribution in Automatic Evaluation by Large Language Models," 2024.

[21] R. Antonello et al., "A Generative Framework to Bridge Data-Driven Models and Scientific Theories in Language Neuroscience," 2024.

[22] Z. Chen, J. Chen, A. K. Singh, and M. Sra, "XplainLLM: A Knowledge-Augmented Dataset for Reliable Grounded Explanations in LLMs," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2024, pp. 7578–7596.

[23] S. H. A. Mahmood, Z. Lu, and M. Yin, "Designing Behavior-Aware AI to Improve the Human-AI Team Performance in AI-Assisted Decision Making," in *Proc. 33rd Int. Joint Conf. Artif. Intell. (IJCAI)*, 2024.

[24] Z. Lu et al., "Mix and Match: Characterizing Heterogeneous Human Behavior in AI-Assisted Decision Making," in *Proc. AAAI Conf. Human Comput. Crowdsourcing (HCOMP)*, vol. 12, 2024.

[25] Z. Li and M. Yin, "Utilizing Human Behavior Modeling to Manipulate Explanations in AI-Assisted Decision Making: The Good, the Bad, and the Scary," in *Proc. 38th Conf. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, BC, Canada, Dec. 2024.

[26] Z. Li, H. Zhu, Z. Lu, Z. Xiao, and M. Yin, "From Text to Trust: Empowering AI-Assisted Decision Making with Adaptive LLM-Powered Analysis," in *Proc. 2025 CHI Conf. Human Factors Comput. Syst.*, Yokohama, Japan, Apr. 2025.

[27] A. Xiong et al., "A Study of Training Strategies on Enhancing Human Detection of Deepfake Images," in *Proc. Int. AAAI Conf. Web Social Media (ICWSM)*, Copenhagen, Denmark, Jun. 2025.

[28] Y. Li et al., "CleanGen: Mitigating Backdoor Attacks for Generation Tasks in Large Language Models," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2024.

[29] S. Momeni, S. Mazumder, Z. Ke, and B. Liu, "In-Context Continual Learning Assisted by an External Continual Learner," in *Proc. 31st Int. Conf. Comput. Linguistics (COLING)*, 2025.

[30] J. Qiu, Z. Ke, and B. Liu, "Continual Learning Using Only Large Language Model Prompting," in *Proc. 31st Int. Conf. Comput. Linguistics (COLING)*, 2025.

[31] G. Kim, C. Xiao, T. Konishi, Z. Ke, and B. Liu, "Open-World Continual Learning: Unifying Novelty Detection and Continual Learning," *Artif. Intell.*, 2024.

[32] K. Lu, J. Chong, M. Lu, and J. Gao, "Enabling Asymptotic Truth Learning in a Social Network," in *Proc. 20th Conf. Web Internet Econ. (WINE)*, Dec. 2024.

[33] F. Uradník et al., "Maximizing Truth Learning in a Social Network Is NP-Hard," in *Proc. 24th Int. Conf. Auton. Agents Multiagent Syst. (AAMAS)*, May 2025.

[34] Z. Xu, F. Jiang, L. Niu, J. Jia, B. Li, and R. Poovendran, "ACE: A Model Poisoning Attack on Contribution Evaluation Methods in Federated Learning," in *Proc. 33rd USENIX Secur. Symp.*, Aug. 2024.

[35] C. Sferrazza, project page. [Online]. Available: https://sferrazza.cc/bot_site

[36] S. Sharma and A. K. Singh, "Dynamic Gradient Influencing for Viral Marketing Using Graph Neural Networks," in *Proc. ACM Web Conf. (WWW)*, 2025, pp. 3982–3993.

[37] C. Deng, J. Gao, K. Lu, F. Luo, H. Sun, and C. Xin, "Neuc-MDS: Non-Euclidean Multidimensional Scaling Through Bilinear Forms," in *Proc. 38th Conf. Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2024.

[38] J. P. Hespanha and K. Camsari, "Markov Chain Monte Carlo for Koopman-Based Optimal Control," *IEEE Control Syst. Lett.*, Jun. 2024, presented at the *63rd IEEE Conf. Decis. Control (CDC)*, Dec. 2024.

[39] K. Zhang, S. Cheng, G. Shen, B. Ribeiro, S. An, P.-Y. Chen, X. Zhang, and N. Li, "CENSOR: Defense Against Gradient Inversion via Orthogonal Subspace Bayesian Sampling," in *Proc. Netw. Distrib. Syst. Secur. Symp. (NDSS)*, 2025.

[40] J. Chenou, T. Fields, and A. Williams, "Collaborative Design Between Intelligent Agents Through Resource Sharing," in *Proc. 11th Int. Conf. Comput. Sci. Comput. Intell. (CSCI)*, Las Vegas, NV, USA, Dec. 2024.

[41] T.-Y. Zhou et al., "Optimal Classification-Based Anomaly Detection with Neural Networks: Theory and Practice," submitted for publication. [Online]. Available: https://arxiv.org/abs/2409.08521

[42] 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.

[43] D. Sahabandu et al., "RL-ARNE: A Reinforcement Learning Algorithm for Computing Average Reward Nash Equilibrium of Nonzero-Sum Stochastic Games," *IEEE Trans. Autom. Control*, vol. 69, no. 11, pp. 7824–7831, Nov. 2024.

[44] S. Moothedath et al., "Stochastic Dynamic Information Flow Tracking Game Using Supervised Learning for Detecting Advanced Persistent Threats," *Automatica*, vol. 159, Art. no. 111353, 2024.

[45] D. Sahabandu et al., "Game of Trojans: Adaptive Adversaries Against Output-Based Trojaned Model Detectors," in *Proc. ACM Asia Conf. Comput. Commun. Secur. (AsiaCCS)*, Jul. 2024.

[46] Y. Wu et al., "Holder Pruning: Localized Pruning for Backdoor Removal in Deep Neural Networks," submitted for publication.

[47] S. Anderson and J. P. Hespanha, "Zero-Sum Turn Games Using Q-Learning: Finite Computation and Security Guarantees," submitted for publication.

[48] T. van Ede et al., "DeepCASE: Semi-Supervised Contextual Analysis of Security Events," in *Proc. 2022 IEEE Symp. Secur. Privacy (SP)*, May 2022, pp. 522–539.

[49] Open Information Security Foundation, "Suricata." [Online]. Available: https://suricata.io/

[50] ACTION AI Institute, "GATE MS Packages," *GitHub*. [Online]. Available: https://github.com/action-ai-institute/gate-ms-packages

[51] UCSB Security Lab, "CVEX," *GitHub*. [Online]. Available: https://github.com/ucsb-seclab/CVEX

[52] Q. Li et al., "LLM-PBE: Assessing Data Privacy in Large Language Models," *Proc. VLDB Endow.*, vol. 17, 2024. [Online]. Available: https://arxiv.org/abs/2408.12787

[53] Y. Yang et al., "SecCodePLT: A Unified Platform for Evaluating the Security of Code GenAI," *arXiv preprint arXiv:2410.11096*, 2024.

[54] C. Wang et al., "Physical ID-Transfer Attacks Against Multi-Object Tracking via Adversarial Trajectory," in *Proc. Annu. Comput. Secur. Appl. Conf. (ACSAC)*, 2024, pp. 1–15.

[55] R. Song et al., "Enhancing LLM-Based Autonomous Driving Agents to Mitigate Perception Attacks," *arXiv preprint arXiv:2409.14488*, 2024.

[56] F. Jiang et al., "ArtPrompt: ASCII Art-Based Jailbreak Attacks Against Aligned LLMs," in *Proc. 62nd Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2024.

[57] M. Mazeika et al., "HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal," in *Proc. 41st Int. Conf. Mach. Learn. (ICML)*, Jul. 2024.

[58] F. Jiang et al., "ChatBug: A Common Vulnerability of Aligned LLMs Induced by Chat Templates," presented at the *NeurIPS 2024 Workshop on Safe Generative AI*, Dec. 2024.

[59] H. Benkraouda, N. Diwan, and G. Wang, "You Can't Judge a Binary by Its Header: Data-Code Separation for Non-Standard ARM Binaries Using Pseudo Labels," in *Proc. 46th IEEE Symp. Secur. Privacy (SP)*, San Francisco, CA, USA, May 2025.

[60] Y. Cheng et al., "CTINEXUS: Leveraging Optimized LLM In-Context Learning for Constructing Cybersecurity Knowledge Graphs Under Data Scarcity," *arXiv preprint arXiv:2410.21060*, 2024.

[61] J. T. Wang et al., "Capturing the Temporal Dependence of Training Data Influence," *arXiv preprint arXiv:2412.09538*, 2024.

[62] C. Katsis, X. Jin, F. Sang, J. Sun, E. Bertino, R. R. Kompella, and A. Kundu, "Graphene: Towards Data-Driven Holistic Security Posture Analysis Using AI-Generated Attack Graphs," in *Proc. 2024 IEEE 10th Int. Conf. Collab. Internet Comput. (CIC)*, 2024, pp. 9–18. [Online]. Available: https://arxiv.org/abs/2312.13119

[63] A. S. Li, A. Iyengar, A. Kundu, and E. Bertino, "Revisiting Concept Drift in Windows Malware Detection: Adaptation to Real Drifted Malware with Minimal Samples," in *Proc. Netw. Distrib. Syst. Secur. Symp. (NDSS)*, 2025.

[64] U. Kaur et al., "RoboCop: A Robust Zero-Day Cyber-Physical Attack Detection Framework for Robots," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2024.

[65] U. Kaur, Z. B. Celik, and R. Voyles, "RoboGuardZ: A Scalable Zero-Shot Framework for Detecting Zero-Day Malware in Robots," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2024.

[66] S. Vasan, H. Aghakhani, S. Ortolani, R. Vasilenko, I. Grishchenko, C. Kruegel, and G. Vigna, "DEEPCAPA: Identifying Malicious Capabilities in Windows Malware," in *Proc. Annu. Comput. Secur. Appl. Conf. (ACSAC)*, Dec. 2024, pp. 826–842.

[67] H. Lu, H. Cai, Y. Liang, A. Bianchi, and Z. B. Celik, "A Progressive Transformer for Unifying Binary Code Embedding and Knowledge Transfer," in *Proc. IEEE Int. Conf. Softw. Anal., Evol. Reeng. (SANER)*, 2025.

[68] IoTDSCreator, "IoTDSCreator," *GitHub*. [Online]. Available: https://github.com/iotdscreator

[69] H. Lee, C. Katsis, A. Lotfi, T. Choi, S. Kim, A. Kundu, and E. Bertino, "IoTDSCreator: A Framework to Create Labeled Datasets for IoT Intrusion Detection Systems," in *Proc. 15th ACM Conf. Data Appl. Secur. Privacy (CODASPY)*, 2025.

[70] K. S. Mubasshir, I. Karim, and E. Bertino, "Gotta Detect 'Em All: Fake Base Station and Multi-Step Attack Detection in Cellular Networks," in *Proc. 34th USENIX Secur. Symp.*, 2025.

[71] J. Allen, Z. Yang, F. Xiao, M. Landen, R. Perdisci, and W. Lee, "WebRR: A Forensic System for Replaying and Investigating Web-Based Attacks in the Modern Web," in *Proc. 33rd USENIX Secur. Symp.*, 2024.

[72] F. Xiao, Z. Su, G. Yang, and W. Lee, "Jasmine: Scale up JavaScript Static Security Analysis with Computation-Based Semantic Explanation," in *Proc. 2024 IEEE Symp. Secur. Privacy (SP)*, 2024.

[73] Amazon Science, "Amazon Trusted AI Challenge." [Online]. Available: https://www.amazon.science/trusted-ai-challenge

[74] Amazon Science, "Team PurpCorn-PLAN." [Online]. Available: https://www.amazon.science/nova-ai-challenge/teams/purpcorn-plan

[75] DARPA, "AI Cyber Challenge." [Online]. Available: https://aicyberchallenge.com/

[76] Shellphish, "Shellphish Support Syndicate." [Online]. Available: https://support.shellphish.net/#about

[77] Team Atlanta, "Team Atlanta." [Online]. Available: https://team-atlanta.github.io/

[78] Computer Science Teachers Association. [Online]. Available: https://csteachers.org/

[79] UC Santa Barbara, "International Capture the Flag (iCTF)." [Online]. Available: https://ictf.cs.ucsb.edu/

[80] pwn.college. [Online]. Available: https://pwn.college/

[81] Berkeley Center for Responsible, Decentralized Intelligence (RDI). [Online]. Available: https://rdi.berkeley.edu/

[82] Berkeley RDI, "Large Language Model Agents MOOC, Fall 2024." [Online]. Available: https://llmagents-learning.org/f24

[83] Berkeley RDI, "LLM Agents Hackathon." [Online]. Available: https://rdi.berkeley.edu/llm-agents-hackathon/

[84] UCTV, "GRIT Talks." [Online]. Available: https://www.uctv.tv/grit/

[85] Center for Security and Emerging Technology, Georgetown University. [Online]. Available: https://cset.georgetown.edu/

[86] Annual Computer Security Applications Conference (ACSAC). [Online]. Available: https://www.acsac.org/

[87] CARTA, University of Miami, "News and Events." [Online]. Available: https://carta.miami.edu/news-and-events/events/index.html

[88] Billington CyberSecurity, "State and Local Cybersecurity Summit." [Online]. Available: https://statelocal.billingtoncybersummit.com/

[89] Z. Xu et al., "Magpie: Alignment Data Synthesis from Scratch by Prompting Aligned LLMs with Nothing." [Online]. Available: https://magpie-align.github.io/

[90] ACTION AI Institute, "MABEL Dataset," *GitHub*. [Online]. Available: https://github.com/action-ai-institute/MABEL-dataset

[91] USC Information Sciences Institute, "SPHERE Project." [Online]. Available: https://sphere-project.net/

[92] EPFL Center for Digital Trust, "About Us." [Online]. Available: https://c4dt.epfl.ch/about-us/

[93] 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.

Appendix A: Glossary and Acronyms

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.

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

AI – Artificial intelligence

AIVO – AI Institutes Virtual Organization

AIxCC – DARPA's AI Cyber Challenge

ASU – Arizona State University

AWS – Amazon Web Services

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

CISA – Cybersecurity and Infrastructure Security Agency

CPS – Cyber-physical Systems

CTI – Cyber Threat Intelligence

CTF – Capture the Flag (security exercise)

CRS – Cyber Reasoning System

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

CVEs – Common Vulnerability and Exposures

DHS – Department of Homeland Security

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

DNNs – Deep Neural Networks

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

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

EPFL – Swiss Federal Institute of Technology Lausanne

GATE– Globally Accessible Test 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 Dongyan Xu, with ACTION

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

MSI – Minority Serving Institution

MSB – Swedish Civil Contingencies Agency

NAIRR – National Artificial Intelligence Research Resource Pilot

NEH – National Endowment for the Humanities

NITRD – Networking and Information Technology Research and Development

NLP – Natural Language Processing

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

R&R – Response and Recovery

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

SAIL – Summit for AI Institute Leadership

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

S&IP or SIP – Strategic and Implementation Plan

SDP – Semidefinite Programming

SoK – Systematization of Knowledge

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).

TI – Threat Intelligence

TTP – Tactics, Techniques, and Procedures

UC – University of California

UCSB – University of California, Santa Barbara

UIUC – University of Illinois, Urbana-Champaign

VPC – Virtual Private Cloud

WiCS – Women in Computer Science

WiCyS – Women in Cybersecurity

Appendix B: Highlights of Significant Achievements and Impact

Highlight 1: Enabling Asymptotic Truth Learning in a Social Network

Outcome/Accomplishment:

Imagine a group of people or AI agents trying to determine the truth about a particular matter. Each person receives a private hint that might be right or wrong and can also observe the decisions of some others who made their choices earlier. The challenge is ensuring that, as more people make decisions, the group as a whole converges on the correct answer.

This research highlight explores how individuals within a network can collectively arrive at the correct decision or belief over time, even when each person has limited and potentially unreliable information.

Key Achievements:

- Fixed or Random Decision Orderings Can Lead to Errors: If individuals make decisions in a given sequence or in a random sequence, especially in sparsely connected networks, the group may not converge on the truth. Early incorrect decisions can mislead others, creating a cascade of errors.

- Structured Decision Sequences Enhance Accuracy: By carefully planning the order in which individuals make decisions—taking into account the network's structure—it's possible to significantly improve the likelihood that the group will collectively arrive at the correct conclusion.

- Applicability to Various Network Types: The researchers demonstrated that their approach works well in different types of networks, including those that resemble real-world social networks, like the Erdős-Rényi and preferential attachment models.

Real-world Impacts:

Reimagine the scenario above as a group of AI agents viewing the network logs, in real time, of a power plant in the United States. At the beginning of a deceptive attack enacted by a malicious actor, can the agents agree on the imminent danger and alert a SoC analyst?

This research provides a framework for enhancing collective decision-making processes, ensuring that groups can effectively discern the truth even when individual members have limited or imperfect information.

Relevant ACTION Research Thrust/Thread:

AI-3: Multi-Agent Collaboration / Resilient multi-agent reasoning

Reference:

See reference [32].

Highlight 2: Learning with contextual information in non-stationary environments

Outcome/Accomplishment:

Anderson and Hespanha's research [42] addresses the challenge of decision-making in environments where conditions change over time. These changes can arise from phenomena such as seasonal trends or, as they explore in this work, adversaries who adapt their strategies over time. The authors introduce a learning algorithm designed to adapt to such non-stationary settings in part by leveraging contextual information available at the time of the decision-making.

Key Achievements:

- Contextual adaptation: The algorithm incorporates contextual information, which are measurements available at the time of a decision, to inform what action to take next. Interestingly, this contextual information could come from the output of other machine learning models that were not designed for changing environments.

- Handling non-stationarity: By continuously updating its understanding based on new data, the algorithm maintains performance even when the underlying environment includes strategic opponents that observe and react to past decisions.

Practical Examples:

The paper demonstrates the algorithm's effectiveness at malware detection, which lacks stationarity because malware designers constantly react to the software that detects it. The proposed algorithm consistently outperforms alternative approaches as malware evolves over time.

Real-world Impacts:

This method helps agents keep learning and improving, even when the world around them is unpredictable – useful for an autonomous cybersecurity system.

Relevant ACTION Research Thrust/Thread:

Thrust AI-4: Strategic Gaming and Tactical Planning / Adversarial modeling and certification.

Reference:

See reference [42].

Highlight 3: Toward Data-driven Holistic Security Posture Analysis using AI-generated Attack Graphs

Outcome/Accomplishment:

Graphene is an automated tool that assesses the security of computer systems by identifying vulnerabilities and mapping out potential attack paths. It utilizes artificial intelligence to analyze vulnerabilities across multiple system layers, including hardware, software, networks, and cryptographic components, and evaluates how weaknesses in one layer can propagate or be exploited to compromise others.

Key Achievements:

- Automated Vulnerability Analysis: Graphene can automatically detect known vulnerabilities in a system by analyzing user-provided information such as device details and software versions.

- AI-Generated Attack Graphs: The system constructs "attack graphs" that illustrate how an attacker might exploit a series of vulnerabilities to compromise a system. This helps in visualizing potential security breaches.

- Risk Scoring Mechanism: Graphene assesses the severity and exploitability of identified vulnerabilities both individually and as part of attack paths, assigning risk scores that help prioritize remediation efforts based on potential impact.

- Holistic Security Assessment: By examining multiple layers of a system and their interconnections, Graphene provides a comprehensive view of an organization's security posture.

Real-world Impacts:

Graphene's approach enables organizations to identify and evaluate security weaknesses before they are exploited proactively. By automating the analysis of complex systems, it significantly reduces the time and expertise required for comprehensive security assessments. This is especially beneficial for large enterprises with intricate infrastructures, where efficient prioritization and resource allocation are critical. Ultimately, Graphene provides security teams with a clear, actionable view of their threat landscape, empowering them to tailor mitigation strategies and reinforce defenses against evolving threats.

Graphene plays a vital role in the knowledge base of ACTION's AI agents, connecting their security assessment abilities (SEC-1) with the Institute's ongoing research in learning and reasoning.

Relevant ACTION Thrust/Thread:

SEC-1: Intelligent Agents for Threat/Vulnerability Assessment / Continual and system-wide vulnerability assessment.

Reference:

See reference [62].

Highlight 4: ArtPrompt: ASCII Art-based Jailbreak Attacks against Aligned LLMs

Outcome/ Accomplishment:

The widespread use of LLMs highlights significant safety concerns. Violations of safety can expose users of LLMs to toxic content or other associated risks. While extensive effort has been made to ensure safety alignment of LLMs, these methods overlook the fact that training/ instruction corpora can be interpreted in multiple ways beyond semantics. We developed a benchmark called Vision-in-Text Challenge and showed that five SOTA LLMs are unable to perceive and respond to queries in the form of ASCII art. We revealed vulnerabilities of LLMs to a novel jailbreak (side-channel) attack called ArtPrompt, which exploits the inability of LLMs to recognize ASCII-art. ArtPrompt was shown to bypass three SOTA defenses against jailbreak attacks.

Key Achievements:

1. **Security Risks:** ArtPrompt demonstrated that malicious actors could exploit this vulnerability to disseminate harmful content, posing significant risks in applications relying on LLMs for content moderation.
2. **Highlighting Limitations in Current Safety Measures:** Existing safety alignment mechanisms in LLMs primarily focus on semantic understanding, neglecting alternative input interpretations like ASCII art.
3. **Need for Enhanced Safety Protocols:** The study emphasizes the necessity for LLM developers to consider non-traditional input forms in their safety mechanisms, prompting a reevaluation of current alignment strategies.
4. **Open-Source**: All ArtPrompt code is public (https://github.com/uw-nsl/ArtPrompt) to enable widespread adoption by the broader community.

Real-World Impact:

The study emphasizes the necessity for LLM developers to consider non-traditional inputs in their safety mechanisms, prompting a reevaluation of current alignment strategies. This work will inspire further investigations into multimodal vulnerabilities in LLMs and the development of more robust defense mechanisms against such attacks. ArtPrompt allows ACTION researchers to focus on agents' reasoning about vulnerabilities, to better understand their root causes and contextualize the security posture of target systems.

Relevant ACTION Thrust/Thread:

SEC-1: Intelligent Agents for Threat/Vulnerability Assessment / Knowledge-guided vulnerability detection.

Reference:

See reference [56].

Highlight 5: DeepCapa: Identifying Malicious Capabilities in Windows Malware

Outcome/ Accomplishment:

In our recent research, we introduce DeepCapa [66], a novel methodology for identifying malicious capabilities present in Windows Malware. DeepCapa operates in two stages. First, it processes activity artifacts (memory dumps) from a malicious program running in a safe, isolated setting, and extracts potential command sequences (API calls) the program might use. This approach enables DeepCapa to analyze both executed and unexecuted malicious code.

In the second phase, DeepCapa uses Artificial Intelligence via neural networks to correlate these command sequences to potentially malicious capabilities.

DeepCapa leverages the state-of-the-art Transformer architecture and is first trained broadly using a technique called masked language modeling (MLM). The AI model is then fine-tuned with examples to detect malicious capabilities. This approach enables DeepCapa to learn from training labels, automatically correlate inputs and outputs, and learn generalized rules.

Another important capability of DeepCapa lies in its ability to localize detected, malicious capabilities.

Real-World Impact:

DeepCapa overcomes a longstanding limitation of malware analysis. First, it overcomes the standard limitations of static and dynamic analysis. Second, DeepCapa uses its AI model to generalize rules for capabilities, enabling it to detect capabilities even when malware uses disguises (i.e. obfuscation, packing) that trick existing rule-based detection methods. Finally, once a capability is detected by DeepCapa, it can then pinpoint where the harmful action is located in the malware's code, providing invaluable insights to malware analysts.

Relevant ACTION Thrust/Thread:

SEC-2: Intelligent Agents for Detection / Novel detection agents for large-scale security incidents.

Reference:

See reference [66].

Highlight 6: MABEL and Heterogeneous Ensemble: Advancing Malware Family Attribution with Scalable AI Models and Granular Analysis

Outcome / Accomplishment

In our research, we introduce a comprehensive framework using Artificial Intelligence that advances malware family attribution through three key contributions:

1. We introduce **MABEL**, the **M**alware **A**nalysis **BE**nchmark for AI and Machine **L**earning [90]. MABEL is a new, expansive benchmark dataset comprising 82,000+ malware samples across 470 malware families, with more than 600 extracted features per sample. Compared to existing malware datasets, MABEL provides greater malware family diversity and deeper feature representation that are actively updated. This dataset allows AI models to more effectively generalize to the complexity and diversity of modern malware.
2. We introduce a novel **heterogeneous classification ensemble** that combines multiple Machine Learning models including Decision Trees, Random Forests, AdaBoost, Support Vector Machine classifiers (SVMs), Naive Bayesian Classifiers (NBCs), and Natural Language Processing (NLP) classification models into a robust classification pipeline applied to malware family attribution. Next, we introduce a ***Classification Arbiter*** that evaluates model performance using diverse metrics (e.g., Macro F1-Score, Precision, Recall, Matthews Correlation Coefficient, Hamming Loss, etc) and selects the most stable prediction for each instance. This ensemble outperforms single-model and homogeneous ensemble baselines across multiple malware families.
3. Feedback and granular evaluation of model performance is crucial for explainability and classification optimization. We developed enhanced classification reporting output that evaluates model performance per malware family and analyzes the contributions of different feature types (e.g., API calls, Strings, opcodes, malware binary metadata). This reporting output identifies which models and features most effective in classifying specific malware families and highlights areas of model uncertainty. This output combined with the heterogeneous classification ensemble and classification arbiter establishes a novel paradigm of building and evaluating heterogeneous classification ensembles that can be applied to different domains.

Impact and Benefits

To our knowledge, this research represents one of the first to amass such a large, feature-rich dataset with malware attributed to malware families and a heterogeneous ensemble that outperforms existing models tested on the MABEL dataset. Our research addresses longstanding limitations in malware family attribution and general AI classification by providing the following key benefits:

1. Our release of MABEL fills a critical dataset gap, enabling academia and industry to benchmark and train AI models on a real-world, diverse, and continuously evolving malware corpus.
2. Our heterogeneous ensemble introduces the concept of efficiently training a large set of diverse models with an arbiter agent that dynamically selects the best classification for each instance in a dataset. This methodology improves classification robustness allowing the arbiter agent to focus on the strengths of each model and mitigates weaknesses present in traditional approaches such as overfitting and poor generalization to rare, underrepresented families.

In total, MABEL, the heterogeneous ensemble, and enhanced reporting represent a major step forward in building explainable, scalable, and more accurate agent-based AI systems for modern cybersecurity challenges.

Relevant ACTION Thrust/Thread:

SEC-3: Intelligent Agents for Attribution / AI-stack tools to assist humans in performing attribution.

Reference:

See references [90, 93].

Appendix C: List of Institute Projects

Thrust AI-1: Learning and Reasoning with Domain Knowledge Research Thread 1: Advanced knowledge representation learning for logical reasoning. Research/Technology -- Direct Support			
Position Title	Name	Department	Institution
Activity Leader	Bo Li	Computer Science	University of Chicago / University of Illinois Urbana-Champaign
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
Thrust AI-1: Learning and Reasoning with Domain Knowledge Research Thread 2: Knowledge discovery through human-in-the-loop and multi-agent systems. Research/Technology -- Direct Support			
Position Title	Name	Department	Institution
Activity Leader	Bo Li	Computer Science	University of Chicago / University of Illinois Urbana-Champaign
Faculty Researcher	Ambuj Singh	Computer Science	University of California, Santa Barbara
Thrust AI-1: Learning and Reasoning with Domain Knowledge Research Thread 3: Veridical secure and private learning with end-to-end quality control. Research/Technology -- Direct Support			
Position Title	Name	Department	Institution
Activity Leader	Bo Li	Computer Science	University of Chicago / University of Illinois Urbana-Champaign
Faculty Researcher	Bin Yu	Statistics	University of California, Berkeley
Faculty Researcher	Sewoong Oh	Computer Science and Engineering	University of Washington
Faculty Researcher	Ambuj Singh	Computer Science	University of California, Santa Barbara

Thrust AI-2: Human-Agent Interaction Research Thread 1: Design adaptive and behavior-aware AI to support human decision-making. Research/Technology -- Direct Support			
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
Thrust AI-2: Human-Agent Interaction Research Thread 2: Efficient solicitation of human knowledge for augmenting AI. Research/Technology -- Direct Support			
Position Title	Name	Department	Institution
Activity Leader	Ming Yin	Computer Science	Purdue University
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
Thrust AI-2: Human-Agent Interaction Research Thread 3: Effective human-agent teaming Research/Technology -- Direct Support			
Position Title	Name	Department	Institution
Activity Leader	Ming Yin	Computer Science	Purdue University
Faculty Researcher	Ambuj Singh	Computer Science	University of California, Santa Barbara
Thrust AI-3: Multi-Agent Collaboration Research Thread 1: Modeling of Multi-Agent Networks. Research/Technology -- Direct Support			
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

Thrust AI-3: Multi-Agent Collaboration
Research Thread 2: Resilient Multi-Agent Reasoning.
Research/Technology -- Direct Support

Position Title	Name	Department	Institution
Activity Leader	Jie Gao	Computer Science	Rutgers University
Faculty Researcher	Radha Poovendran	Electrical and Computer Engineering	University of Washington
Faculty Researcher	Pieter Abbeel	Electrical Engineering and Computer Science	University of California, Berkeley
Faculty Researcher	Ambuj Singh	Computer Science	University of California, Santa Barbara

Thrust AI-3: Multi-Agent Collaboration
Research Thread 3: Robust and Adversary-Aware Reinforcement Learning.
Research/Technology -- Direct Support

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	Wenke Lee	Computer Science	Georgia Institute of Technology

Thrust AI-4: Strategic Gaming and Tactical Planning
Research Thread 1: Integration of Knowledge and Information Flow Graphs for Enhancing Cyber AI Games.
Research/Technology -- Direct Support

Position Title	Name	Department	Institution
Activity Leader	Radha Poovendran	Electrical and Computer Engineering	University of Washington

Faculty Researcher	Wenke Lee	Computer Science	Georgia Institute of Technology
Thrust AI-4: Strategic Gaming and Tactical Planning Research Thread 2: Adversarial Modeling and Certification. Research/Technology -- Direct Support			
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
Thrust AI-4: Strategic Gaming and Tactical Planning Research Thread 3: Deception Modeling Approaches. Research/Technology -- Direct Support			
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
Thrust SEC-1: Intelligent Agents for Threat/Vulnerability Assessment Research Thread 1: Constructing Knowledge Base for Threat Assessment. Research/Technology -- Direct Support			
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
Thrust SEC-1: Intelligent Agents for Threat/Vulnerability Assessment Research Thread 2: Knowledge-guided Vulnerability Detection. Research/Technology -- Direct Support			
Position Title	Name	Department	Institution
Activity Leader	Gang Wang	Computer Science	University of Illinois at Urbana-Champaign

Faculty Researcher	Giovanni Vigna	Computer Science	University of California, Santa Barbara
Faculty Researcher	Christopher Kruegel	Computer Science	University of California, Santa Barbara
Faculty Researcher	Bo Li	Computer Science	University of Chicago / University of Illinois Urbana-Champaign
Faculty Researcher	Dawn Song	Electrical Engineering and Computer Science	University of California, Berkeley
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
Thrust SEC-1: Intelligent Agents for Threat/Vulnerability Assessment Research Thread 3: Continual and System-wide Vulnerability Assessment. Research/Technology -- Direct Support			
Position Title	Name	Department	Institution
Activity Leader	Gang Wang	Computer Science	University of Illinois at Urbana-Champaign
Faculty Researcher	Elisa Bertino	Computer Science	Purdue University
Thrust SEC-2: Intelligent Agents for Detection Research Thread 1: More Robust Agents for Detection Research/Technology -- Direct Support			
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
Thrust SEC-2: Intelligent Agents for Detection Research Thread 2: Handling Concept Drift Research/Technology -- Direct Support			
Position Title	Name	Department	Institution
Activity Leader	David Evans	Computer Science	University of Virginia

Faculty Researcher	Elisa Bertino	Computer Science	Purdue University
Faculty Researcher	Sewoong Oh	Computer Science and Engineering	University of Washington

Thrust SEC-2: Intelligent Agents for Detection Research Thread 3: Novel Detection Agents for Large-Scale Security Incidents Research/Technology -- Direct Support			
Position Title	Name	Department	Institution
Activity Leader	David Evans	Computer Science	University of Virginia
Faculty Researcher	Berkay Celik	Computer Science	Purdue University
Faculty Researcher	Giovanni Vigna	Computer Science	University of California, Santa Barbara
Faculty Researcher	Christopher Kruegel	Computer Science	University of California, Santa Barbara

Thrust SEC-3: Intelligent Agents for Attribution Research Thread 3: AI-Stack Tools to Assist Humans in Performing Attribution Research/Technology -- Direct Support			
Position Title	Name	Department	Institution
Activity Leader	Nick Feamster	Computer Science	University of Chicago
Faculty Researcher	Dongyan Xu	Computer Science	Purdue University

Thrust SEC-4: Intelligent Agents for Response and Recovery Research Thread 1: Identification of rare events and new attack plan Research/Technology -- Direct Support			
Position Title	Name	Department	Institution
Activity Leader	Elisa Bertino	Computer Science	Purdue University
Faculty Researcher	Wenke Lee	Computer Science	Georgia Institute of Technology

Thrust SEC-4: Intelligent Agents for Response and Recovery Research Thread 2: Automated and Robust Attack Recovery Research/Technology -- Direct Support			
Position Title	Name	Department	Institution
Activity Leader	Elisa Bertino	Computer Science	Purdue University

Faculty Researcher	Berkay Celik	Computer Science	Purdue University
Faculty Researcher	Wenke Lee	Computer Science	Georgia Institute of Technology
Thrust SEC-4: Intelligent Agents for Response and Recovery Research Thread 3: Efficient and Adaptive Response Strategies Research/Technology -- Direct Support			
Position Title	Name	Department	Institution
Activity Leader	Elisa Bertino	Computer Science	Purdue University
Faculty Researcher	Dongyan Xu	Computer Science	Purdue University
Faculty Researcher	Berkay Celik	Computer Science	Purdue University
Globally Accessible Test Environment (GATE) Research/Technology & Broadening Participation -- Direct Support			
Position Title	Name	Department	Institution
Activity Leader	Berkay Celik	Computer Science	Purdue University
Software Engineer	Noah Spahn	Institute Staff	University of California, Santa Barbara
Activity Leader	Wenke Lee	Computer Science	Georgia Institute of Technology
Faculty Researcher	Dongyan Xu	Computer Science	Purdue University
Faculty Researcher	Giovanni Vigna	Computer Science	University of California, Santa Barbara
Postdoctoral Scholar	Abdulellah Alsaheel	Computer Science	Purdue University
Faculty Researcher	Bo Li	Computer Science	University of Chicago / University of Illinois Urbana-Champaign
Faculty Researcher	David Evans	Computer Science	University of Virginia
Graduate Student	Mathew Lau	Computer Science	Georgia Institute of Technology
Graduate Student	Doguhan Yeke	Computer Science	Purdue University
Graduate Student	Jizhou Chen	Computer Science	Georgia Institute of Technology

Graduate Student	Yupeng Yang	Computer Science	Georgia Institute of Technology
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
Faculty Researcher	Gang Wang	Computer Science	University of Illinois, Urbana-Champaign
Faculty Researcher	Yung-hsiang Lu	Electrical and Computer Engineering	Purdue University
Undergraduate Student	Gabriel Pizarro	Computer Science	University of California, Santa Barbara
Undergraduate Student	Tianle Yu	Computer Science	University of California, Santa Barbara
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	University of California, Santa Barbara
Faculty Researcher	Christopher Kruegel	Computer Science	University of California, Santa Barbara
Faculty Researcher	Wenke Lee	Computer Science	Georgia Institute of Technology
Faculty Researcher	Dongyan Xu	Computer Science	Purdue University
Software Engineer	Noah Spahn	Institute Staff	University of California, Santa Barbara
Graduate Student	Laboni Sarker	Computer Science	University of California, Santa Barbara
High School Outreach **Education and Workforce Development & Broadening Participation-- Direct Support**			
Position Title	Name	Department	Institution

Activity Leader	Giovanni Vigna	Computer Science	University of California, Santa Barbara
Managing Director	Tim Robinson	Institute Staff	University of California, Santa Barbara
Graduate Student	Stijn Pletinckx	Computer Science	University of California, Santa Barbara
Postdoctoral Scholar	Ilya Grishchenko	Computer Science	University of California, Santa Barbara
Graduate Student	Laboni Sarker	Computer Science	University of California, Santa Barbara

Summer Research Experience

Education and Workforce Development -- Direct Support

Position Title	Name	Department	Institution
Activity Leader	Tim Robinson	Institute Staff	University of California, Santa Barbara
Faculty Researcher	Giovanni Vigna	Computer Science	University of California, Santa Barbara
Software Engineer	Noah Spahn	Institute Staff	University of California, Santa Barbara
Activity Leader	Lukas Dresel	Computer Science	University of California, Santa Barbara
Postdoctoral Scholar	Yigitcan Kaya	Computer Science	University of California, Santa Barbara
Undergraduate Student Intern	Anton Landerer	Computer Science	University of California, Santa Barbara
Undergraduate Student Intern	Michelle Zimmerman	Computer Science	University of California, Santa Barbara
Undergraduate Student Intern	Luca Paliska	Computer Science	De Anza College, California
Undergraduate Student Intern	Ilda Martinez	Computer Science	Cal State University, San Bernardino
Graduate Student Intern	Swati Saxena	Computer Science	Cal State University, San Bernardino
Undergraduate Student Intern	Jonathan Aguilar	Computer Science	Cal State University, San Bernardino
Undergraduate Student Intern	Grace Jin	Computer Science	Cornell University

Graduate Student Intern	Marco Cerrato	Computer Science	Cal State University, San Bernardino
Graduate Student Intern	Weiheng Bai	Computer Science	University of Minnesota
Undergraduate Student Intern	Evelyn Gutierrez	Computer Science	Cal State University, San Bernardino
Undergraduate Student Intern	Austin Chan	Computer Science	University of California, Berkeley
Graduate Student Intern	Guy Munn	Computer Science	Norfolk State University
Undergraduate Student Intern	Erik Feng	Computer Science	University of California, Santa Barbara
Undergraduate Student Intern	Vincent Huynh	Computer Science	American River College, California
Undergraduate Student Intern	Filip Uradník	Computer Science	Charles University, Czech Republic
Undergraduate Student Intern	Julia Krizanova	Computer Science	Charles University, Czech Republic
Undergraduate Student Intern	Amanda Wang	Computer Science	Princeton University
Undergraduate Student Intern	Rhett Olson	Computer Science	University of Minnesota
Undergraduate Student Intern	Wendon Doswell	Computer Science	Norfolk State University
Undergraduate Student Intern	Clay Jones	Computer Science	Norfolk State University
Undergraduate Student Intern	Korey Bailey	Computer Science	Norfolk State University
Faculty Researcher	Tonya Fields	Computer Science	Norfolk State University
Faculty Researcher	Jules Chenou	Computer Science	Norfolk State University
Undergraduate Student Intern	Jasmine McDuffie	Computer Science	Norfolk State University
Graduate Student Intern	Timothy Agboada	Computer Science	Norfolk State University

Appendix D: Project Updates

Summary slides follow.

Thrust AI-1: Learning and Reasoning with Domain Knowledge

Research thread: Advanced knowledge representation learning for logical reasoning

Challenge

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

Overarching Goal

Integrate statistical learning with domain knowledge and reasoning ability for AI agents.

Planned goals, June 2024 - May 2025

- Build generic learning-reasoning frameworks for different applications (e.g., intrusion detection and malware detection).
- Explore different efficient knowledge representation learning approaches for large language models.
- Design domain-specific languages to represent domain knowledge of diverse cybersecurity tasks, which can be easily adapted to new tasks.
- Build a large retrieval-augmented generative model to integrate knowledge into the prediction and reasoning processes at scale.

Proposed Research

- Develop a robust sensing-reasoning pipeline, integrating statistical machine learning (e.g., DNNs, RLs) with logic-based reasoning.
- Leverage domain knowledge to enhance inference beyond statistical predictions.
- Cybersecurity domains are complex, interconnected systems – inherently uncertain, partial evidence.

Achievements this period

- Automated red teaming approach against code generation models.
- Method to enhance LLMs' ability to handle compositional tasks through decomposition into smaller problems.
- System to ensure AI agents do not have incentive to deceive humans.
- Method for agents to incrementally learn sequences of tasks consisting of multiple classes of graphs.

Bo Li[1,2], Dawn Song[1,3], Stuart Russell[1,3], Bing Liu[1,4]

[1] Faculty, [2] University of Chicago, Computer Science, [3] UC Berkeley, Computer Science, [4] University of Illinois, Chicago, Computer Science

Thrust AI-1: Learning and Reasoning with Domain Knowledge

Research thread: Knowledge discovery through human-in-the-loop and multi-agent systems

Challenge

Current knowledge graph-based approaches lack the scalability and lifelong learning ability to extract domain knowledge to empower AI models with trackable logic reasoning ability.

Overarching Goal

Design and develop scalable and lifelong continuous knowledge extraction for AI agents.

Planned goals, June 2024 - May 2025

- Map team decision-making & human feedback into formal knowledge representations.
- Use human feedback to align agent behavior with real-world.
- Develop a scalable system for cyber threat knowledge from diverse sources using advanced NLP and labeling techniques.
- Generate synthetic data from simulations, enhanced with diffusion and generative models, to enhance model learning.

Proposed Research

- Develop framework for automated, high-fidelity cyber threat knowledge acquisition using AI techniques like NLP and lifelong learning.
- System constructs and continuously updates threat knowledge graphs.
- Continual learning-based method using labeled and unlabeled cyber threat intelligence.

Achievements this period

- Exploration of counterfactual learning in the context of human guidance on generative models.
- Investigation into how well LLMs are able to solve problems on graphs.

Ambuj Singh[1,2]

[1] Faculty, [2] UC Santa Barbara, Computer Science

Thrust AI-1: Learning and Reasoning with Domain Knowledge

Research thread: Veridical secure and private learning with end-to-end quality control

Challenge

Current learning approaches lack the end-to-end quality control ability to estimate and enhance the stability and explainability of trained agents.

Overarching Goal

Provide veridical (truthful) learning and reasoning for AI agents.

Planned goals, June 2024 - May 2025

- Use malware data to validate and evaluate the framework to provide further feedback and improve the safety of the end-to-end pipeline.
- Optimize prompts to periodically check the reliability of LLMs to ensure the quality of LLMs.
- Design and develop advanced reward functions to help integrate human feedback to guide the knowledge acquisition from LLMs.

Proposed Research

- Incorporate enhanced Predictability, Computability, and Stability (PCS) framework for quality control.
- Support reliable learning and reasoning for cybersecurity agents using both real-world and simulated attack data.
- Enable comparison between predicted and actual outcomes for stability and quality assessments of the pipeline.
- Facilitate knowledge discovery via hypothesis generation.

Achievements this period

- Novel approach for anomaly detection on graph data that addresses limitations of purely statistical data.
- REC (Rate, Explain and Cite) framework – a systematic method for automatic evaluation using LLMs.
- Sample- and time-efficient differentially private algorithm for ordinary least squares.

Bo Li[1,2], Bin Yu[1,3], Sewoong Oh[1,4]

[1] Faculty, [2] University of Chicago, Computer Science, [3] UC Berkeley, Computer Science, [4] University of Washington, Seattle, Computer Science and Engineering

Thrust AI-2: Human-Agent Interaction

Research thread: Design adaptive and behavior-aware AI to support human decision-making

Challenge

Current human-AI systems are weak in supporting humans' responses to agents' recommendations and actions.

Overarching Goal

Develop principled methods for effective human-agent interaction with agent-in-the-loop decision-making.

Planned goals, June 2024 - May 2025

- Create a basic model where AI assists an individual decision-maker, to understand how humans prioritize AI-generated intelligence.
- Incorporate prospect theory, influence dynamics, and Bayesian learning to account for biases, external influences, and risk evaluation among decision-makers.
- Capture how individuals in a group collectively make decisions based on risks and available information.

Proposed Research

- Promote humans' calibrated trust in agents.
- Model human behavior in agent-in-the-loop decision-making.
- Design agents that are aware of and adaptive to human behavior.

Achievements this period

- Computational models of decision makers' reactions to AI recommendations and explanations to improve AI-human team performance.
- Intervention strategies to help human users generalize knowledge from agents.
- Work on LLMs supporting humans triage alerts from intrusion detection systems.

Ming Yin[1,2], Gang Wang [1,3], David Wagner [1,4]

[1] Faculty, [2] Purdue University, Computer Science, [3] University of Illinois, Urbana-Champaign, Computer Science, [4] University of California, Berkeley, Computer Science

Thrust AI-2: Human-Agent Interaction

Research thread: Efficient solicitation of human knowledge for augmenting AI

Challenge

Lack of principled ways to solicit high-level human intelligence (i.e., beyond labels) in a scalable and reliable manner to augment autonomous agents.

Overarching Goal

Develop principled methods for effective human-agent interaction with learning and reasoning with humans-in-the-loop.

Planned goals, June 2024 - May 2025

- Seek human feedback on agent actions and behavior from an individual.
- Use this feedback to adjust agent training (e.g., reward adjustment) and improve agent performance.
- Explore soliciting feedback from a group of stakeholders with different perspectives and conflicting interests.
- Find ways to balance their diverse feedback and update the agent in a way that accounts for these conflicting objectives.

Proposed Research

- Investigate methods to solicit domain knowledge from humans.
- Investigate methods to solicit feedback from humans.
- Investigate the potential of implicitly soliciting humans' domain knowledge and intuition from human behavioral data.

Achievements this period

- Human+ML systems for detecting jailbreak attacks against LLMs.
- Mitigation strategies for safety alignment in LLMs.
- Novel method that achieves the upper bound accuracy of continual learning, interactive with human users.

David Wagner [1,2], Radha Poovendran [1,3], Bing Liu [1,4]

[1] Faculty, [2] UC Berkeley, Computer Science, [3] University of Washington, Electrical and Computer Engineering, [4] University of Illinois, Chicago, Computer Science

Thrust AI-2: Human-Agent Interaction

Research thread: Effective human-agent teaming

Challenge

Lack of principled ways to conduct task allocation and information exchange within a team composed of both humans and agents to optimize the effectiveness of the team.

Overarching Goal

Develop principled methods for effective human-agent interaction with human-agent teaming.
Two types of human-agent teaming: Division of labor and collaboration with interdependence.

Planned goals, June 2024 - May 2025

- Build user-friendly interfaces powered by LLMs to facilitate communication and negotiation between human and AI agents.
- Support multiple team structures: equal teams of humans and AI agents; and human-led teams.
- Enhance the interfaces with algorithms to recommend the most effective division of labor.

Proposed Research

- Optimize task allocation in division of labor between humans and agents.
- Enhance communication strategies between humans and agents when tasks influence one another.
- Model mutual understanding: enable two-way communication through strategic information “push” and “pull” mechanisms.

Achievements this period

- Collaborative project initiated to explore how LLMs can serve as facilitators within human-AI groups to improve group decision making outcomes.
- Delphi method-like style to ensure equal participation of group members.

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

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

Thrust AI-3: Multi-Agent Collaboration

Research thread: Modeling of multi-agent networks

Challenge

Current multi-agent frameworks lack an analytical model to assess the aggregate capability of a network of agents.

Overarching Goal

Develop a multi-agent collaboration framework, enabled by research in the modeling of multi-agent networks.

Planned goals, June 2024 - May 2025

- Design and analyze techniques for info exchange among agents.
- Analyze overall capabilities of a set of agents.
- Pursue applications to real-world settings. Examine scalability.
- Focus on multi-agent learning. When and how to support successful learning and avoid information cascade (aka herding).
- Improve the resource efficiency of data processing on edge devices that provide sensor data to agents.

Proposed Research

- Model a dynamic multi-agent network where agents share knowledge to detect cyber threats.
- Use a contagion-inspired framework for timely dissemination of discoveries, enabling coordinated responses.
- Develop protocols balancing knowledge spread, conflict resolution, and resilient reasoning across the agent network.

Achievements this period

- Research on graph topology to inform communication and decision making abilities of collaborating agents.
- Representations for networks of interacting dynamical agents using Koopman operator theory and Markov Chain Monte Carlo algorithms.
- Research on privacy concerns in federated learning to collaboratively train a neural network on a global server.

Jie Gao [1,3], João Hespanha [1,4], Ninghui Li [1,5], Jules Chenou [2, 6], Tonya Fields [1,6]

[1] Faculty, [2] Researcher, [3] Rutgers, Computer Science, [4] UC Santa Barbara, Electrical and Computer Engineering, [5] Purdue, Computer Science, [6] Norfolk State University, Cybersecurity

Thrust AI-3: Multi-Agent Collaboration

Research thread: Resilient multi-agent reasoning

Challenge

Current multi-agent reasoning approaches do not consider agent heterogeneity and asynchrony, or absence of knowledge.

Overarching Goal

Develop a multi-agent collaboration framework, enabled by research in the modeling of multi-agent reasoning.

Planned goals, June 2024 - May 2025

- Model the limitations of multi-agent reasoning using different kinds of data.
- Study the techniques for resiliency in reasoning:
  - Analysis of concept drift
  - Analysis of optimality
  - Enabling federated learning (sharing models vs. data)
  - Poisoning agents and malicious information passing
- Investigate applications to real-world scenarios.

Proposed Research

- Enable efficient, targeted knowledge queries in multi-agent networks using a knowledge type system and distributed trust-based discovery.
- Support reasoning under uncertainty.
- Accelerate decisions through speculative reasoning and expert queries, even with incomplete or delayed information.

Achievements this period

- Research on understanding reasoning in a networked, multi-agent scenario with uncertainty in observations.
- Model poisoning attack on contribution evaluation methods in federated learning.
- Research on how a multi-agent structuring of high degree of freedom control problems can speed up learning and make inference more efficient.

Jie Gao [1,2], Radha Poovendran [1,3], Pieter Abbeel[1,4]

[1] Faculty, [2] Rutgers University, Computer Science, [3] University of Washington, Electrical and Computer Engineering, [4] UC Berkeley, Electrical Engineering and Computer Science

Thrust AI-3: Multi-Agent Collaboration

Research thread: Robust and adversary-aware reinforcement learning

Challenge

Current multi-agent learning approaches underestimate the existence of adversarial agents and conflicting views of the world among agents.

Overarching Goal

Develop a multi-agent collaboration framework, enabled by research in the modeling of multi-agent adversary-aware learning.

Planned goals, June 2024 - May 2025

- Design reinforcement learning algorithms for the security domain, which requires the analysis of a game played between defender and adversary, with models for action costs.
- We will develop formal approaches to learn attackers' intent and capabilities in adversarial settings, specifically where opponents actively want to hide their intent/capabilities.

Proposed Research

- Develop resilient multi-agent policies using adversary-aware training to counter stealthy, adaptive attacks in adversarial environments.
- Use multi-agent reinforcement learning (MARL) with adversary-embedded datasets to learn robust policies under sparse and misleading feedback.

Achievements this period

- First theoretical guarantees of unsupervised neural network-based anomaly detectors and empirical insights on how to design them well.

Xiaoming Huo [1,2], Wenke Lee [1,3]

[1] Faculty, [2] Georgia Institute of Technology, Statistics, [3] Georgia Institute of Technology, Computer Science

Thrust AI-4: Strategic Gaming and Tactical Planning

Research thread: Integration of Knowledge and IFG for enhancing cyber AI games

Challenge

Current game-theoretical approaches mainly focus on defending against anticipated attacks, with little reasoning about the dynamic nature of the attackers' behavior.

Overarching Goal

Establish a composable layer of defense for a security system that realizes AI agents' strategies and tactics, enabled by research in foundation of strategy games.

Planned goals, June 2024 - May 2025

- Develop graph structures integrating KGs and IFGs to capture interactions between AI-enabled adversaries and AI-driven system defenses in the Cyber AI game.
  - The IFG will be constructed using system log data and will provide post-mortem knowledge about the impact of an attack.
  - As the IFG evolves, snippets of the IFG will contain information about events related to system execution.

Proposed Research

- Model cyber-AI interactions as dynamic, non-cooperative games between intelligent agents and stealthy adversaries.
- Develop detection and deception games to prioritize inspection points and mislead adversaries.
- Create non-equilibrium-seeking cyber games and "games-of-games" to model evolving strategies.

Achievements this period

- New stochastic game models evolving on information graphs characterizing the interactions between dynamic information flow tracking (DIFT) and detect advanced persistent threats (APTs).
- Work on the design of defense mechanisms against backdoor attacks on deep neural networks (DNNs)

Radha Poovendran [1,2], Wenke Lee[1,3]

[1] Faculty, [2] University of Washington, Electrical and Computer Engineering, [3] Georgia Institute of Technology, Computer Science

Thrust AI-4: Strategic Gaming and Tactical Planning

Research thread: Adversarial modeling and certification

Challenge

Currently, there is no approach that provides a worst-case performance bound (minimal guarantee) for a detection strategy in non-cooperative dynamic, non-equilibrium-seeking games.

Overarching Goal

Establish a composable layer of defense for a security system that realizes AI agents' strategies and tactics, enabled by research in adversary/threat modeling and defense function certification.

Planned goals, June 2024 - May 2025

- Model interactions between two players as an adversarial reinforcement learning problem
- Study adversarial RL in the latent space. How should players interact to maximize their payoffs?
  - First, examine when state and action spaces are discrete
  - Subsequently investigate more complex continuous state and action spaces
- Develop LP and SDP relaxations to enable certifiable bounds.

Proposed Research

- Combine security game strategies with statistical robustness certification to train resilient agents and certify performance under constrained adversarial attacks.
- Develop scalable solutions for large-scale cyber games using LP/SDP relaxations and exploit submodularity to manage exponential action space growth.

Achievements this period

- Work on repeated decision making with contextual or side information, where a decision maker needs to select an action to minimize a prescribed cost with feedback from measurements that provide context.
- Algorithm inspired by no-regret log-linear learning that uses Boltzmann distributions to generate stochastic policies

João Hespanha [1,2]

[1] Faculty, [2] UC Santa Barbara, Electrical and Computer Engineering

Thrust AI-4: Strategic Gaming and Tactical Planning

Research thread: Deception modeling approaches

Challenge

Current detection systems are incapable of dealing with fast and high-volume information flows.

Overarching Goal

Establish a composable layer of defense for a security system that realizes AI agents' strategies and tactics, enabled by research in the dynamic planning of agent tactics.

Planned goals, June 2024 - May 2025

- Develop a hypergame framework to model the strategic and deceptive interactions between AI-enabled adversaries and the AI defense in Cyber AI games.
- Exploit the hierarchical and coupled structures inherent to hypergames and develop techniques to synthesize deception strategies as emergent solutions.
- Investigate and identify classes of hypergames for which it will be possible to determine the best response at each stage and provide performance bounds.

Proposed Research

- Model cyber games with information asymmetry to derive defense strategies like deception, moving target defense, and obfuscation as emergent outcomes.
- Develop multi-round deception strategies using signaling games, local deception planning, and lifelong learning integrated with risk assessment.
- Counter adversarial probing and misinformation.

Achievements this period

- Work towards development of Reinforcement Learning (RL) algorithms that maximize the quality of RL policies obtained with finite computation and the design of stopping criteria that can guarantee minimum levels of performance.
- Pure saddle-point state-feedback policies can be constructed using a suitable form of Q-learning.

João Hespanha [1,2]

[1] Faculty, [2] UC Santa Barbara, Electrical and Computer Engineering

AI Institute for Agent-based Cyber Threat Intelligence and Operation

Globally Accessible Testing Environment (GATE)

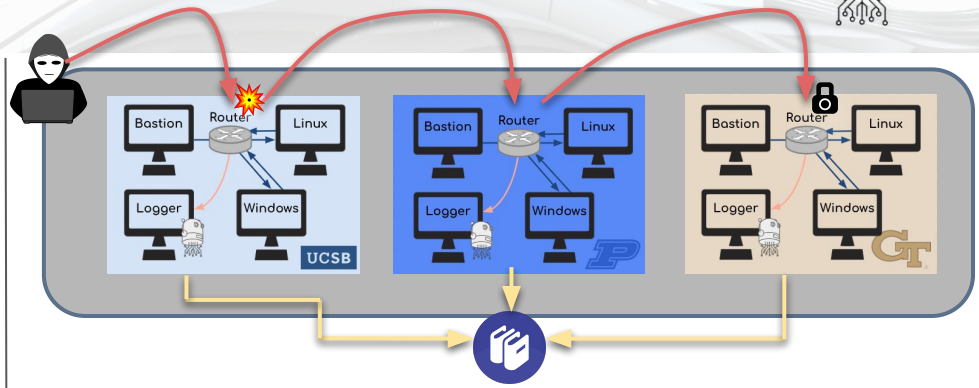
Overarching Goal

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

- Use of GATE for novel, hands-on educational activities during Summer 2024
- Inclusion of NPCs (e.g. CMU GHOSTS) to generate realistic background traffic
- Man-in-the-Middle (MitM) proxy to add value to traffic logs
- Simplified end user experience from new command line interface (CLI)
- Cloud-based cyber-range now template-based for changeability
- Pre-configured VPN from researcher's machine to the cloud
- Demonstrated sophisticated attack scenario leveraging CVE-2024-4577
- "Starter-kit" helps explain interplay between ACTION Server and AI Stack
- Expanded collaborative team across Institute and beyond – e.g. USC SPHERE

Z. Berkay Celik[1,3], Noah Spahn[2,5], Dongyan Xu[1,3], Wenke Lee[1,4], Giovanni Vigna [1,5] et al.

[1] Faculty, [2] Software Engineer, [3] Purdue University, Computer Science, [4] Georgia Institute of Technology, Computer Science, [5] UC Santa Barbara, Computer Science

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

Research thread: Constructing Knowledge Base for Threat Assessment

Challenge

Current approaches do not support the integration of knowledge about both threats and targets in a cohesive fashion.

Overarching Goal

Develop a holistic framework, **focused on system-threat knowledge extraction**, that extends vulnerability analysis from individual system components (e.g., programs) to a composite system (e.g., a critical infrastructure), revealing assets, attack surfaces, and potential vulnerabilities.

Proposed Research

- Create an integrated, dynamically evolving knowledge base to capture both system-specific and cyber threat knowledge.
- Use extracted information from human experts and system data, forming a structured knowledge graph.
- Develop knowledge agent for system-wide threat and vulnerability assessments.

Planned goals, June 2024 - May 2025

- Survey available data sources, gather datasets, and construct preliminary knowledge base.
- Enhance the knowledge base, in collaboration with Thrust AI-1.
- Explore the vulnerability knowledge base beyond the cyber domain to the physical domain, by modeling, simulation, and knowledge derivation for cyber-physical system (CPS) security by leveraging domain-specific languages and retrieval-augmented generative models from Thrust AI-1.

Achievements this period

- Loadstar combines machine learning with domain knowledge for binary analysis.
- Work on parsing unstructured and semi-structured log files at scale, enabling intelligent search to support human security analysts.
- CTINexus leverages LLMs for constructing cybersecurity knowledge graphs.
- Study on the temporal impacts of data influence on ML.

Gang Wang [1,2], David Wagner [1,3], Dawn Song [1,3]

[1] Faculty, [2] University of Illinois, Urbana-Champaign, Computer Science, [3] UC Berkeley, Computer Science,

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

Research thread: Knowledge-guided vulnerability detection

Challenge

Current vulnerability discovery techniques identify vulnerabilities without semantic/contextual system knowledge.

Overarching Goal

Develop a holistic framework, **focused on unit component assessment**, that extends vulnerability analysis from individual system components (e.g., programs) to a composite system (e.g., a critical infrastructure), revealing assets, attack surfaces, and potential vulnerabilities.

Proposed Research

- Create unit assessment agents to assess system components using system-threat knowledge and human collaboration. Agents will aim to identify vulnerabilities by leveraging control models and other system knowledge.
- Division of labor between agents and humans will evolve over time. Agents gradually assume more complex tasks.

Planned goals, June 2024 - May 2025

- Focus on logic reasoning about the vulnerabilities, to better understand the root causes and contextualize the security posture of the target system.
- Develop “CVE-X” that produces the containers to run PoC (proof-of-concept) exploits to reproduce and verify vulnerabilities.
- Generalize the knowledge base with the goal of discovering new vulnerabilities based on the known ones.
- Address the challenge of reasoning about low-frequency/rare vulnerabilities.

Achievements this period

- CVEX (eXecutable CVEs) framework to enable reproducible exploitation of Common Vulnerabilities and Exposures (CVEs).
- LLM-PBE toolkit analyzes privacy vulnerabilities using diverse attack and defense strategies.
- RedCode introduces a platform to evaluate the safety of AI-driven code agents, focusing on risks in code execution and generation.
- Exploration of vulnerabilities in LLMs through ArtPrompt and ChatBug.
- Improvements to Ghidra, state-of-the-art decompiler.

Giovanni Vigna [1,2], Christopher Kruegel [1,2], Bo Li [1,3], Dawn Song [1,4], Berkay Celik [1,5], Radha Poovendran [1,6], Dongyan Xu [1,5]

[1] Faculty, [2] UC Santa Barbara, Computer Science, [3] U Chicago, Computer Science, [4] UC Berkeley, Computer Science, [5] Purdue, Computer Science, [6] U Washington, Electrical Engineering

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

Research thread: Continual and system-wide vulnerability assessment

Challenge

Current threat assessment approaches tend to be opportunistic and lack a whole-system view of the target and its security posture.

Overarching Goal

Develop a holistic framework, **focused on system-wide assessment**, that extends vulnerability analysis from individual system components (e.g., programs) to a composite system (e.g., a critical infrastructure), revealing assets, attack surfaces, and potential vulnerabilities.

Proposed Research

- Enable agents to perform proactive security assessments.
- Enable agents to uncover multi-step attack paths, even between physically separate networks, by linking vulnerabilities in small components to broader, system-wide risks.
- Allow collaboration and knowledge acquisition among agents.

Planned goals, June 2024 - May 2025

- Integrate generative models and paradigms to systematically integrate the domain knowledge for cybersecurity with prediction and reasoning to enhance vulnerability analysis and discovery for CPS.
- Focus on system-wide vulnerability assessments by modeling the relationships of vulnerabilities across different system entities (software, hosts, and networks).

Achievements this period

- Graphene, a system to provide a detailed analysis of security posture of computing infrastructure. Collaboration with Cisco.

Elisa Bertino [1,2]

[1] Faculty, [2] Purdue University, Computer Science

Thrust SEC-2: Intelligent Agents for Detection

Research thread: More robust agents for detection

Challenge

Current learning approaches are brittle and degrade when facing incomplete, low-fidelity training data.
High-quality, context-rich training data for AI agents is often scarce or incomplete, leading to false positives and missed attacks during reasoning.

Overarching Goal

Develop approaches for the detection of complex, multi-step attacks, that provide robust training of detection agents.

Planned goals, June 2024 - May 2025

- Explore transfer learning to train detection agents. Assign labels to data from a similar target domain with few or no labels.
  - Incorporate adversarial domain adaptation (DA).
  - Generate domain-independent representation (DIRep) of the data from different domains.
- Begin experiments to improve the accuracy of our adversarial domain adaptation (DA).

Proposed Research

- Develop multi-source and privacy-preserving adversarial domain adaptation (using GANs and DP-GANs) to synthesize realistic training data from disparate, privacy-sensitive sources.
- Use unsupervised clustering to help human analysts efficiently improve label quality and identify new attack classes from sparsely or poorly labeled datasets.

Achievements this period

- Framework developed to automate binary analysis to produce standardized datasets ready for AI/ML modeling.
- MABEL dataset submitted to NAIRR.
- Novel attack surface on LLMs identified, using BPE tokenizers.

Dongyan Xu [1,2], Sewoong Oh [1,3]

[1] Faculty, [2] Purdue University, Computer Science, [3] University of Washington, Computer Science and Engineering

Thrust SEC-2: Intelligent Agents for Detection

Research thread: Handling concept drift

Challenge

Current AI approaches are inadequate in the cybersecurity domain due to adaptive and evasive adversaries.

Overarching Goal

Develop approaches for the detection of complex, multi-step attacks, that provide handling of attacker adaptation, threat evolution and evasion (e.g. concept drift).

Planned goals, June 2024 - May 2025

- Study techniques to detect the fact that a model is experiencing a loss of detection accuracy due to concept drift.
- Build knowledge-based techniques to diagnose the root cause of this drift. Perform multi-dimensional analysis.
- Extend the ability to detect concept drift and missed attacks with the help of human experts.
    - Build effective ways in which agents can explain their decisions.
    - Develop means to communicate these explanations to human experts.

Proposed Research

- Address concept drift with models that detect and label drifting samples while maintaining previously learned knowledge to prevent catastrophic forgetting.
- Enhance robustness by leveraging prior research on adversarial examples.
- Fuses system and threat knowledge to enrich ML feature spaces and improve context-aware, resilient threat detection.

Achievements this period

- Novel, graph-based malware detection system using adversarial training from a small number of drifted malware samples outperforms state-of-the-art.
- Differentially private (DP) machine learning pipelines that can better handle distribution shifts.

Elisa Bertino [1,2], Sewoong Oh [1,3]

[1] Faculty, [2] Purdue University, Computer Science, [3] University of Washington, Computer Science and Engineering

Thrust SEC-2: Intelligent Agents for Detection

Research thread: Novel detection agents for large-scale security incidents

Challenge

Current detection methods are unreliable and result in low accuracy in uncertain contexts.

Overarching Goal

Develop approaches for the detection of complex, multi-step attacks, that provide novel detection approaches incorporating uncertainty measures and multi-agent collaboration, especially for organizations.

Planned goals, June 2024 - May 2025

- Work on novel AI stack-based models to detect malware and the capabilities of malicious code.
- Develop an agent that can identify potentially malicious capabilities in malware using a formal representation of the MITRE ATT&CK knowledge base.
- Explore byte-level data augmentation (ByteDA) to improve the performance of our binary models on malware classification tasks.

Proposed Research

- Develop detection agents that use novel input signals, infer system invariants, and apply advanced anomaly detection and behavior-based models.
- Reduce false positives and prioritize meaningful alerts by correlating signals across multiple intelligent agents.
- Implement multi-agent collaboration using graph-based models and distributed reasoning to construct attack narratives.

Achievements this period

- Robocop to detect zero-days against CPS.
- RoboGuardZ to detect zero-days in robots and multi-sensor autonomous car controllers.
- DEEPCAPA outperforms state-of-the-art in malware detection.

Z. Berkay Celik [1,2], Giovanni Vigna [1,3], Christopher Kruegel [1,3]

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

Thrust SEC-3: Intelligent Agents for Attribution

Research thread: AI-stack tools to assist humans in performing attribution

Challenge

Current attribution techniques cannot directly handle obfuscation & evasive approaches by sophisticated adversaries.

Overarching Goal

Develop a novel attribution approach that focuses on attack scenario attribution by leveraging Large Language Models (LLMs) and Knowledge Graphs (KGs).

Planned goals, June 2024 - May 2025

- Use LLMs combined with KGs to provide a more intuitive and efficient natural language-based user interface to humans.
- Explore ways to construct a new set of fine-tuned, KG-enhanced LLMs for attribution tasks.
- Leverage LLMs to develop effective/optimal prompting strategies for different KG tasks, such as construction and question answering.

Proposed Research

- Deploy Attribution Agents to assess MMAS simulation outputs against real forensic event streams.
- Facilitate human-agent collaboration using work from Thrust AI-2 to iteratively refine simulations by incorporating analyst feedback, attacker goals, and evasion tactics (Thrust AI-4).
- Generate meta-signatures that capture consistent MIA-action patterns across simulations.

Achievements this period

- Natural Language Processing (NLP) models to enhance malware family attribution and characterization.

Dongyan Xu [1,2]

[1] Faculty, [2] Purdue University, Computer Science

Thrust SEC-4: Intelligent Agents for Response and Recovery

Research thread: Identification of rare events and new attack plan

Challenge

Current response approaches are not effective against dynamically evolving multi-step attack scenarios and new/rare events.

Overarching Goal

Develop novel approaches to response and recovery that focus on analysis and containment of multi-step attacks.

Planned goals, June 2024 - May 2025

- Develop a multi-detector framework, for response purposes, to investigate anomalies caused by each attack step and analyze a log system that records system activity at various levels of detail.
- Evaluate the robustness of the multi-detector approach against white-box attacks and consequently develop the proper re-training mechanisms, as a response to the earlier attacks.
- Develop attack correlation and causality analysis algorithms for analyzing the events identified by the multi-detector system.

Proposed Research

- Develop detection and response agents that detect anomalies, trace infection paths, and match event sequences to known attack graphs to predict and contain next attacks.
- Integrate lifelong learning and self-supervised techniques to enable agents to recognize, assess, and adapt in real time.
- Design response mechanisms that empower analysts to guide agents in timely and effective countermeasures.

Achievements this period

- Jasmine, a computation-based semantic explanation (CSE) approach, can effectively resolve complex semantics and discover hidden vulnerabilities.

Wenke Lee [1,2]

[1] Faculty, [2] Georgia Institute of Technology, Computer Science

Thrust SEC-4: Intelligent Agents for Response and Recovery

Research thread: Automated and robust attack recovery

Challenge

Current approaches are not robust against sophisticated adversarial manipulations.

Overarching Goal

Develop novel approaches to response and recovery that focus on robust and scalable forensics against manipulation attacks.

Planned goals, June 2024 - May 2025

- Leverage examples of the existing “security actuators” to enable intelligent agents to learn the models of response actions, disrupt an attack, and improve the security posture of a target.
- Create examples of training-time and inference-time attacks,
- Develop an LLM-based approach to generate variants of these attacks.
- Extend our LLM-based automation approach to source and binary code.

Proposed Research

- Develop forensic analysis tools to enable scalable post-attack traceback for both training-time and inference-time attacks.
- Generate attack-specific datasets and knowledge through forensic investigations to improve attack response strategies.
- Enhance agents' robustness by integrating root cause analysis into training workflows, strengthening model resilience.

Achievements this period

- ProTST, progressive transfer learning framework for binary code embedding.
- IoTDSCreator and FSBDetector systems for predicting and defending against future intrusions.
- WEBRR for post-attack forensic analysis.

Berkay Celik [1,2], Elisa Bertino [1,2], Wenke Lee [1,3]

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

Thrust SEC-4: Intelligent Agents for Response and Recovery

Research thread: Efficient and adaptive response strategies

Challenge

Current response and recovery approaches are ad-hoc and specific to attack scenarios.

Overarching Goal

Develop novel approaches to response and recovery that focus on planning and adaptation of response and recovery.

Planned goals, June 2024 - May 2025

- Design efficient heuristic islanding schemes that have diminishing efficiency returns or submodular property with respect to the number of islands, powered by our AI stack's formal reasoning capability, to provide provable guarantees of safety with provable bounds on the selected response strategy.
- Begin to extend solution to consider sophisticated attacks that impact multiple networks/sectors.

Proposed Research

- Integrate cost-aware, game-theoretic models (AI-4) for optimization while enabling human feedback (AI-2).
- Leverage agent teaming from AI-3 to form collaborative defense collectives.
- Train and test R&R strategies guided by methods from SEC-3 and SEC-2 to ensure resilient, context-aware recovery planning.

Achievements this period

- SecXP framework, a self-explainable neural network architecture designed for security, addresses limitations of existing explainable AI. Integrates human-interpretable explanations directly into model's architecture, rather than rely on post-hoc analysis.

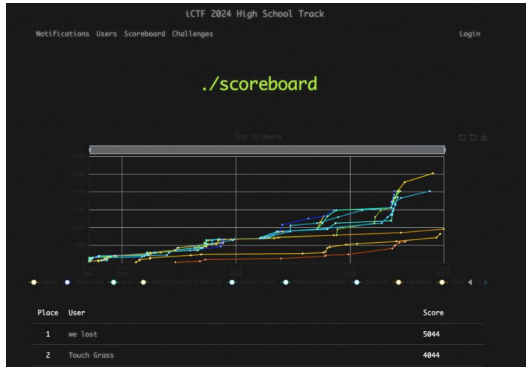
Dongyan Xu [1,2], Z. Berkay Celik [1,2]

[1] Faculty, [2] Purdue University, Computer Science

Education: International Capture the Flag (iCTF) Security + AI Training through CS Problem Solving

Project Overview

- Revision of successful iCTF, educational security exercises. CTF are fun and approachable
- Jeopardy-style CTF incorporating AI, both as human-agent teaming and AI as the subject of security challenges
- Separate tracks for high school and undergraduate
- Year 2 goals: grow iCTF, sponsored by ACTION
- iCTF can be re-packaged for education and workforce development (potential for industry collaboration)



Pictured: iCTF 2024 Scoreboard

Challenges

- Many HS and college students are unprepared for the demands of CTF, or they are hesitant to participate
- Advertising the event efficiently – school only starts two months prior
- 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 challenge design for different levels of students

Achievements this period

- 60% growth in HS iCTF participation
- Greater number of challenges in both HS and Undergraduate tracks
- 12 training videos/walkthroughs on YouTube
- Collaboration with Women in Computer Science (WiCS) to promote the event beyond ACTION campuses
- Data collected from all participation to improve next year's project
- Participants overwhelmingly believed the experience to be beneficial
- iCTF connects ACTION with high schools for increased outreach

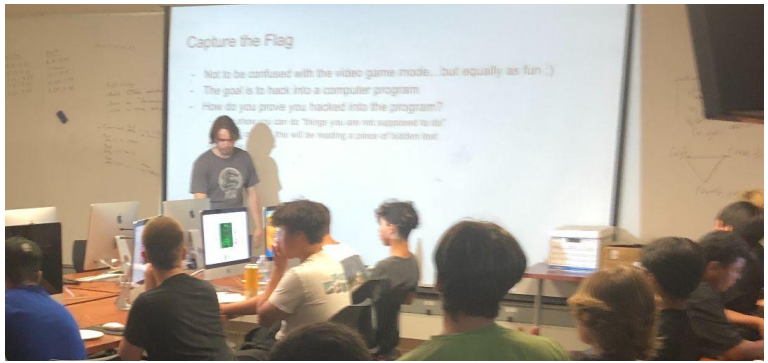
Giovanni Vigna [1,4], Christopher Kruegel [1,4], Wenke Lee [1,5], Dongyan Xu [1,6], Noah Spahn [3,4], Laboni Sarker [2,4]

[1] Faculty, [2] Graduate Student, [3] Software Engineer, [4] UC Santa Barbara, Computer Science, [5] Georgia Institute of Technology, Computer Science, [6] Purdue University, 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, September 2024.

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 scheduled for June 21 at UCSB

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] Managing Director, [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



Pictured: 2024 Summer Cohort at UC Santa Barbara.

Challenges

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

Solutions

- Establish recruiting partnerships with underrepresented student groups/organizations across ACTION, as well as minority serving institutions (MSI)
- Mentor training through collaboration with established campus programs on ACTION campuses

Achievements this period

- 18 undergraduates recruited. 5 graduate students from 10 different colleges and universities
- Multiple academic paper submissions from results; one conference publication
- Students work in teams of 3 or 4 alongside senior-level ACTION graduate student, postdoc, or faculty
- Projects at UCSB, Rutgers, and Norfolk State
- Secure application portal deployed – more than 100 applications received and safely stored
- Help from ACTION Student Advisory Council to select students
- Cohort recruited for 2025, ready to start in June

Tim Robinson [1,6], Giovanni Vigna [2,6], Noah Spahn 3[,6], Lukas Dresel [4,6], Yigitcan Kaya [5,6], ACTION Student Advisory Council [4,5,7]

[1] Managing Director,[2] Faculty, [3] Software Engineer, [4] Graduate Student, [5] Postdoctoral Scholar, [6] UC Santa Barbara, Computer Science, [7] All ACTION Campuses.