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 1 Annual Report

Submitted May 24, 2024
Approved for Public Release

Cooperative agreement: 2229876



Table of Contents

Project Summary **3**
Participant Tables **4**
TABLE 1 – PRINCIPAL ORGANIZATIONS OF THE INSTITUTE 4
TABLE 2 - LIST OF THE LEADERSHIP TEAM 5
TABLE 3 - LIST OF MAJOR INSTITUTE ACTIVITY PARTICIPANTS 7
TABLE 4 - LIST OF OTHER COLLABORATORS AND PARTNERS 10
TABLE 5 - ADVISORY BOARDS 10
Executive Summary **12**
Narrative: **16**
1.0 Vision 16
2.0 Strategic Plan 17
3.0 Essential Functions of the Institute 18
3.1 Research 18
3.1.1 Foundational AI research 21
3.1.2 Other Institute research 22
3.1.3 Benefits of use-inspired research 22
3.2 Nurture and grow the next generation of talent 23
3.2.1. Pre-college education 23
3.2.2 University, College, and Community College Education 25
3.2.3 Workforce development 26
3.2.4 Informal Learning and Education 27
3.3 Broadening participation and diversity 27
3.3.1 International Capture the Flag 27
3.3.2 High School Outreach 28
3.3.3 Summer Research Experiences 29
3.4 Multidisciplinary integration and community building 29
3.4.1 Seminars 29
3.4.2 Campus Events 30
3.4.3 Other Meetings and Communication 30
3.5 Multi-organizational synergies/achievements 31
3.6 Knowledge transfer 32
3.6.1 Publications 32
3.6.2 Presentations 32
3.6.3 Other Meetings Related to Knowledge Sharing 33
3.7 Impact of the Institute as a nexus point for collaborative efforts 33
3.7.1 Media 33
3.7.2 New grants 34
3.7.3 Other collaborative efforts 34
4.0 Evaluation and Assessment 34

5.0 Infrastructure and Management 35
5.1. Configuration and Leadership 35
5.1.1 Organizational Chart 35
5.1.2 Ethics Training 36
5.2 Management Systems and Reporting 36
5.2.1 Management Systems 36
5.2.1.1 Role of advisory boards. 36
5.2.1.2 Project management 36
5.2.1.3 The role of mentors 37
5.2.2 Conflicts of Interest 37
5.3 Resources and Organizational Equipment 37
References **39**
Appendix A: Glossary and Acronyms 41
Appendix B: Project Updates 42

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, these solutions 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:
The proposed research will revolutionize the process of securing and defending complex, large-scale systems, driven by new foundational AI concepts, models, and methods. 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. As a result, 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 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
Senior Personnel	Bing Liu	Computer Science	University of Illinois at Chicago
Senior Personnel	Nick Feamster	Computer Science	University of Chicago
Senior Personnel	Radha Poovendran	Electrical and Computer 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	Sewoong Oh	Computer Science and Engineering	University of Washington
Senior Personnel	Tonya Fields	Computer Science	Norfolk State University
Senior Personnel	Mary Ann Hoppa	Computer Science	Norfolk State University
Managing Director	Tim Robinson	Institute Staff	University of California, Santa Barbara
Software Engineer	Noah Spahn	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	Bo Li	Computer Science	University of Chicago
Faculty Researcher	Bin Yu	Statistics	University of California, Berkeley
Faculty Researcher	Dawn Song	Electrical Engineering and Computer Science	University of California, Berkeley
Faculty Researcher	Stuart Russell	Electrical Engineering and Computer Science	University of California, Berkeley
Faculty Researcher	Bing Liu	Computer Science	University of Illinois at Chicago

Faculty Researcher	Ambuj Singh	Computer Science	University of California, Santa Barbara
Faculty Researcher	Ming Yin	Computer Science	Purdue University
Faculty Researcher	Radha Poovendran	Electrical and Computer 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	Bing Liu	Computer Science	University of Illinois at Chicago
Faculty Researcher	Ambuj Singh	Computer Science	University of California, Santa Barbara
Faculty Researcher	Ninghui Li	Computer Science	Purdue University
Faculty Researcher	Radha Poovendran	Electrical and Computer Engineering	University of Washington
Faculty Researcher	Mary Ann Hoppa	Computer Science	Norfolk State University
Faculty Researcher	David Wagner	Electrical Engineering and Computer Science	University of California, Berkeley
Faculty Researcher	Gang Wang	Computer Science	University of Illinois at Urbana-Champaign
Faculty Researcher	Bo Li	Computer Science	University of Chicago
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	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
Faculty Researcher	Yung-Hsiang Lu	Electrical and Computer Engineering	Purdue University
Faculty Researcher	João Hespanha	Electrical and Computer Engineering	University of California, Santa Barbara
Thrust AI-4: Strategic Gaming and Tactical Planning			
Position Title	Name	Department	Institution
Activity Leader	Radha Poovendran	Electrical and Computer Engineering	University of Washington
Faculty Researcher	João Hespanha	Electrical and Computer Engineering	University of California, Santa Barbara
Faculty Researcher	Bo Li	Computer Science	University of Chicago
Faculty Researcher	Giovanni Vigna	Computer Science	University of California, Santa Barbara
Faculty Researcher	Wenke Lee	Computer Science	Georgia Institute of Technology

Thrust SEC-1: Intelligent Agents for Threat/Vulnerability Assessment			
Position Title	Name	Department	Institution
Activity Leader	Gang Wang	Computer Science	University of Illinois at Urbana-Champaign
Faculty Researcher	Dongyan Xu	Computer Science	Purdue University
Faculty Researcher	Wenke Lee	Computer Science	Georgia Institute of Technology
Faculty Researcher	David Evans	Computer Science	University of Virginia
Faculty Researcher	Giovanni Vigna	Computer Science	University of California, Santa Barbara
Faculty Researcher	Dawn Song	Electrical Engineering and Computer Science	University of California, Berkeley
Faculty Researcher	Christopher Kruegel	Computer Science	University of California, Santa Barbara
Faculty Researcher	Ninghui Li	Computer Science	Purdue University
Faculty Researcher	Berkay Celik	Computer Science	Purdue University
Thrust SEC-2: Intelligent Agents for Detection			
Position Title	Name	Department	Institution
Activity Leader	Christopher Kruegel	Computer Science	University of California, Santa Barbara
Faculty Researcher	Dongyan Xu	Computer Science	Purdue University
Faculty Researcher	David Evans	Computer Science	University of Virginia
Faculty Researcher	Dawn Song	Electrical Engineering and Computer Science	University of California, Berkeley
Faculty Researcher	Ninghui Li	Computer Science	Purdue University
Faculty Researcher	Nick Feamster	Computer Science	University of Chicago
Faculty Researcher	Elisa Bertino	Computer Science	Purdue University
Faculty Researcher	Gang Wang	Computer Science	University of Illinois at Urbana-Champaign
Faculty Researcher	Giovanni Vigna	Computer Science	University of California, Santa Barbara
Faculty Researcher	Radha Poovendran	Electrical and Computer Engineering	University of Washington
Faculty Researcher	Berkay Celik	Computer Science	Purdue University
Faculty Researcher	Sewoong Oh	Computer Science and Engineering	University of Washington
Faculty Researcher	João Hespanha	Electrical and Computer Engineering	University of California, Santa Barbara
Thrust SEC-3: Intelligent Agents for Attribution			
Position Title	Name	Department	Institution
Activity Leader	Giovanni Vigna	Computer Science	University of California, Santa Barbara
Faculty Researcher	David Evans	Computer Science	University of Virginia
Faculty Researcher	Bo Li	Computer Science	University of Chicago
Faculty Researcher	Wenke Lee	Computer Science	Georgia Institute of Technology
Faculty Researcher	Dawn Song	Electrical Engineering and Computer Science	University of California, Berkeley
Faculty Researcher	Tonya Fields	Computer Science	Norfolk State University
Faculty Researcher	Dongyan Xu	Computer Science	Purdue University

Faculty Researcher	Radha Poovendran	Electrical and Computer Engineering	University of Washington
Faculty Researcher	Christopher Kruegel	Computer Science	University of California, Santa Barbara
Faculty Researcher	Yung-Hsiang Lu	Electrical and Computer Engineering	Purdue University
Faculty Researcher	Sewoong Oh	Computer Science and Engineering	University of Washington
Thrust SEC-4: Intelligent Agents for Response and Recovery			
Position Title	Name	Department	Institution
Activity Leader	Wenke Lee	Computer Science	Georgia Institute of Technology
Faculty Researcher	Nick Feamster	Computer Science	University of Chicago
Faculty Researcher	Bo Li	Computer Science	University of Chicago
Faculty Researcher	Elisa Bertino	Computer Science	Purdue University
Faculty Researcher	Giovanni Vigna	Computer Science	University of California, Santa Barbara
Faculty Researcher	David Wagner	Electrical Engineering and Computer Science	University of California, Berkeley

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
Rick Echevarria	VP	Security Sales	Intel Corporation
Tom Gillis	SVP/GM	Security	Cisco
Latanya Sweeney	Professor of the Practice of Government and Technology	Harvard 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 Kwon	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
Ozgur Ozmen	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
Anshuman Suri	PhD Student	Computer Science	University of Virginia

Executive Summary

Narrative

ACTION's first program year can be seen as a ramping up of activity and an opportunity for the many researchers to familiarize themselves with one another and explore opportunities for collaboration. The first year provided ACTION faculty researchers with a chance to see where their current interests could fit into the ACTION research plan.

Regarding research, ACTION faculty have published more than 35 manuscripts in top tier conference publications and journals, and they have at least that many more to be published by the end of the summer. There are three summer conferences where ACTION research is well represented. The 2024 Computer Vision and Pattern Recognition (CVPR) conference in June in Seattle, WA will see four separate ACTION papers. The 2024 International Conference on Machine Learning (ICML) in Vienna, Austria in July will see another 16 papers; and this achievement is on the heels of the Twelfth International Conference on Learning Representations (ICLR), which also took place in Vienna. ACTION researchers presented six research papers there. A number of papers will also be presented at USENIX Security in August.

Given that each of ACTION's eight research thrusts has three research threads, there are 24 primary research activities ongoing within the Institute. The development of the GATE infrastructure marks research activity number 25. Table 3 lists the faculty participating in each research thrust, and Appendix B summarizes progress on each project.

Most of the research produced this year is foundational, supporting the knowledge base for the AI stack that will inform our intelligent security agents. Researchers also made progress in agent interaction and reasoning, as well as in adversarial modeling. Also, the security researchers involved in ACTION reveal ample evidence of AI vulnerabilities, including opportunities for LLM jailbreaks and attack misattribution. These results will contribute to a robust knowledge base for threat assessment and novel reasoning about how best to defend a network, or a series of networks.

This annual report highlights one security result in particular, both for its promise in monitoring industrial control systems and for its deployment within the collaborative GATE environment. Researchers at Purdue present "SAIN: Improving ICS Attack Detection Sensitivity via State-Aware Invariants," which will be published at the USENIX Security Conference in August 2024. When the Las Palomas Power Plant in the Institute's modeled, fictional city of New Esperanza suffered a sophisticated cyber attack, SAIN not only detected the attack and alerted human analysts to take further action, but it also maintained the power plant in a suboptimal safe state.

ACTION activities over the past year have not been centered entirely on research. The Institute made significant progress along an ambitious plan to grow the next generation of AI+security enthusiasts at all levels of education (K-12, Undergraduate, Graduate, and Postdoctoral). ACTION launched online security exercises, which are inspired by human-agent interaction and AI technology, for both high school and undergraduate students. With their basis in fundamental problem-solving, these CTF exercises have the potential to reach new populations of students that typical computer science courses might not, especially at the K-12 level. ACTION researchers have already begun to use these CTFs in high school education and plan to expand this work over the next academic year. ACTION received a Faculty Outreach Grant from the Chancellor's Outreach Advisory Board at UC Santa Barbara and has plans to maintain a presence at local schools for boots-on-the-ground interaction with young students, who may benefit most from cybersecurity training.

ACTION has also developed an immersive summer research program at UC Santa Barbara with the goal of increasing diversity in Computer Science and providing research opportunities to talented and motivated undergraduate and graduate students. The Institute's inaugural summer program will kick off in June 2024 with eight undergraduate and four graduate students teaming up with senior PhD students and postdocs on 8-week projects. These 12 students were selected from a pool of more than 100 applicants, primarily from Minority Serving Institutions across the country.

Regarding community building, this year ACTION launched a seminar series that connects the community through distinguished lectures and student-run presentations. With this seminar series and a number of invited talks from faculty researchers, the Institute is building its role as a nexus of academic and industry interaction.

The structure put in place by the initial SIP still remains. Both the External Advisory Board (EAB) and the Student Advisory Council (SAC) have assembled into well-recognized units within the organization. ACTION's Executive Committee meets weekly and communication among the 30+ institute personnel is constantly improving.

Achievements in the Reporting Period:

i. Foundational AI research
   a. 35 Publications, with 35 forthcoming by October, and 12 more either submitted or under revision.
   b. Collaborative GATE testbed development.
   c. Successful employment of SAIN anomaly detection methodology on GATE.

ii. Other Institute research
   a. Advising on government regulation of AI.

iii. Benefits of use-inspired research
   a. Transfer of continuous learning to other domains.
   b. Contributions to DARPA AIxCC challenge

iv. University and college education
   a. iCTF for undergraduate and graduate students (Dec 2 - Dec 8).
   b. New AI course planning.
   c. Launch of summer research experience.

v. Pre-college education
   a. iCTF for high school students (Dec 2 - Dec 8).
   b. Summer 2023 and 2024 Internship
   c. High school outreach and UCSB Chancellor's Faculty Outreach Grant

vi. Workforce development
   a. Continued plans to generate online curricula for use of CTF challenges for cybersecurity training.

vii. Broadening participation and diversity
   a. Impacts of iCTF and summer research experience on diversity.
   b. UCSB Chancellor's Outreach Advisory Board grant for high school outreach.

viii. Multidisciplinary integration and community building
   a. ACTION Seminar Series.
   b. GATE development.
   c. Broadening ACTION's community to include Philosophy/Ethics specialists.

ix. Multi-organizational synergies/achievements
   a. ACTION faculty co-organize a session on Statistical Methods in Cybersecurity at the Joint Statistical Meeting (JSM 2024).
   b. ACTION faculty lead the Trojan Detection Challenge at NeurIPS 2023.
   c. All-hands-on-deck call for iCTF challenge creation leads to participation from across ACTION Institute.

x. Knowledge transfer
   a. ACTION faculty presentations
   b. DARPA site visit team receives presentation on ACTION activities.
   c. Research collaborations between ACTION researchers and industry.

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. Development of a Community of Practice at UCSB includes ACTION.

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. By fostering collaboration among the member organizations and with external industry partners, the ACTION Institute will become a nexus for both the AI and cybersecurity communities.

2.0 Strategic Plan

The ACTION Institute's Strategic and Implementation Plan (SIP) defines the managerial structure of the ACTION Institute, and outlines the eight research thrusts to be undertaken in order to achieve the proposed goals. Each thrust includes a discussion of challenges and suggests plans for overcoming these challenges. Each thrust is then divided into three interdependent research "threads" that will run through the first five years of the Institute. Each thread is 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, primarily because of the integrated nature of the research. The SIP also describes, at a high level, the various education and outreach activities the Institute will employ to make an impact in the community and future workforce.

ACTION researchers provided an initial version of a SIP in August 2023, after holding a strategic retreat at UC Santa Barbara. In May 2024, at the end of our first program year, we provided a revised SIP containing more details on how research thrusts and threads would be integrated and how achievements would be measured. The revised SIP also includes additional details and justifications for our chosen strategies for broadening participation and increasing diversity, equity and inclusion. Our outreach plans have not changed over the past year; rather, we have had a chance to define how we will implement those plans. Finally, the revised SIP includes more details concerning the role of the advisory committees, as well as general management of the organization.

Throughout the revised strategic plan, 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.

- Thrust AI-1: 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-2: How can humans interact meaningfully with intelligent security agents to provide required domain knowledge and optimize the assignment of tasks?
- Thrust AI-3: How can one assess the capabilities of multiple interacting intelligent agents under adversarial influence and in a fast-changing environment?
- Thrust AI-4: What kind of game-theoretical approaches can effectively model adversarial behavior while providing guarantees on detection effectiveness and the ability to deceive sophisticated attackers?

As highlighted in the SIP, 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 support the identification of adversarial collaboration and will 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 SIP places additional focus on the Globally Accessible Test Environment (GATE) infrastructure, which will be a collaborative effort among the entire ACTION team. At the end of each program year, there will be 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 SIP on at least 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 first program year, ACTION progressed research along the eight research thrusts as described in the Institute's SIP. With each research thrust being broken in three separate threads, and counting GATE as a separate endeavor, there are 25 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 each of the 25 projects

either informs or builds upon other activities, often spanning multiple thrusts or through execution in the GATE environment.

Table 3 lists the faculty participating in each research thrust. Appendix B provides details on specific activities over the last 12 months 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 these co-investigators.

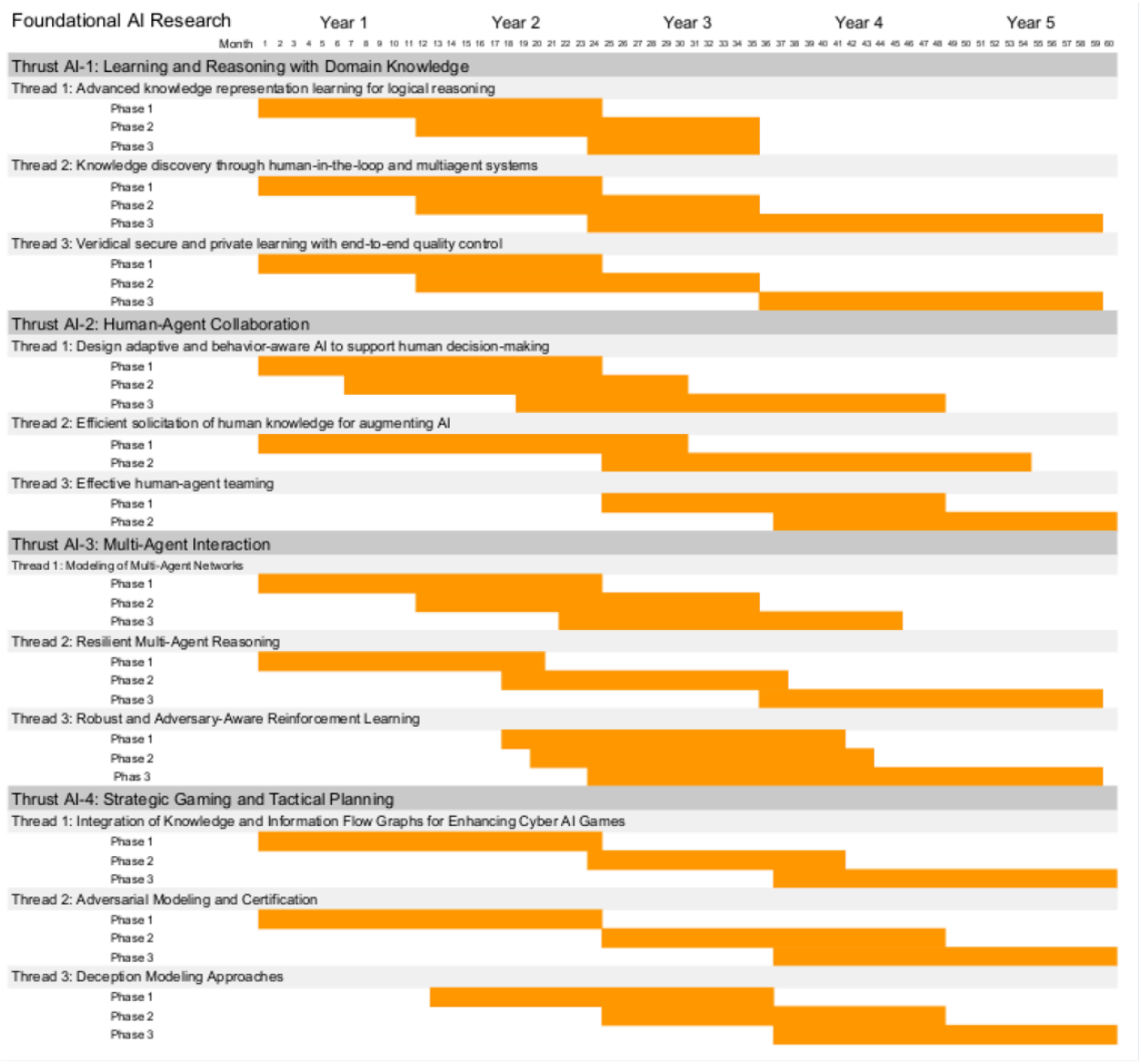


Figure 1: Timeline of AI research activities

We have not encountered any significant delays or unforeseen challenges to our planned research progress. Nor are there any new major research directions. The timelines proposed in our SIP (see Fig. 1 and Fig. 2) remain accurate as of our first year of activity. Three of the activities (Thrust AI-2-Thread 3, Thrust AI-3-Thread 3, and Thrust AI-4-Thread 3) are all set to begin after Year 1 and will have minimal updates here.

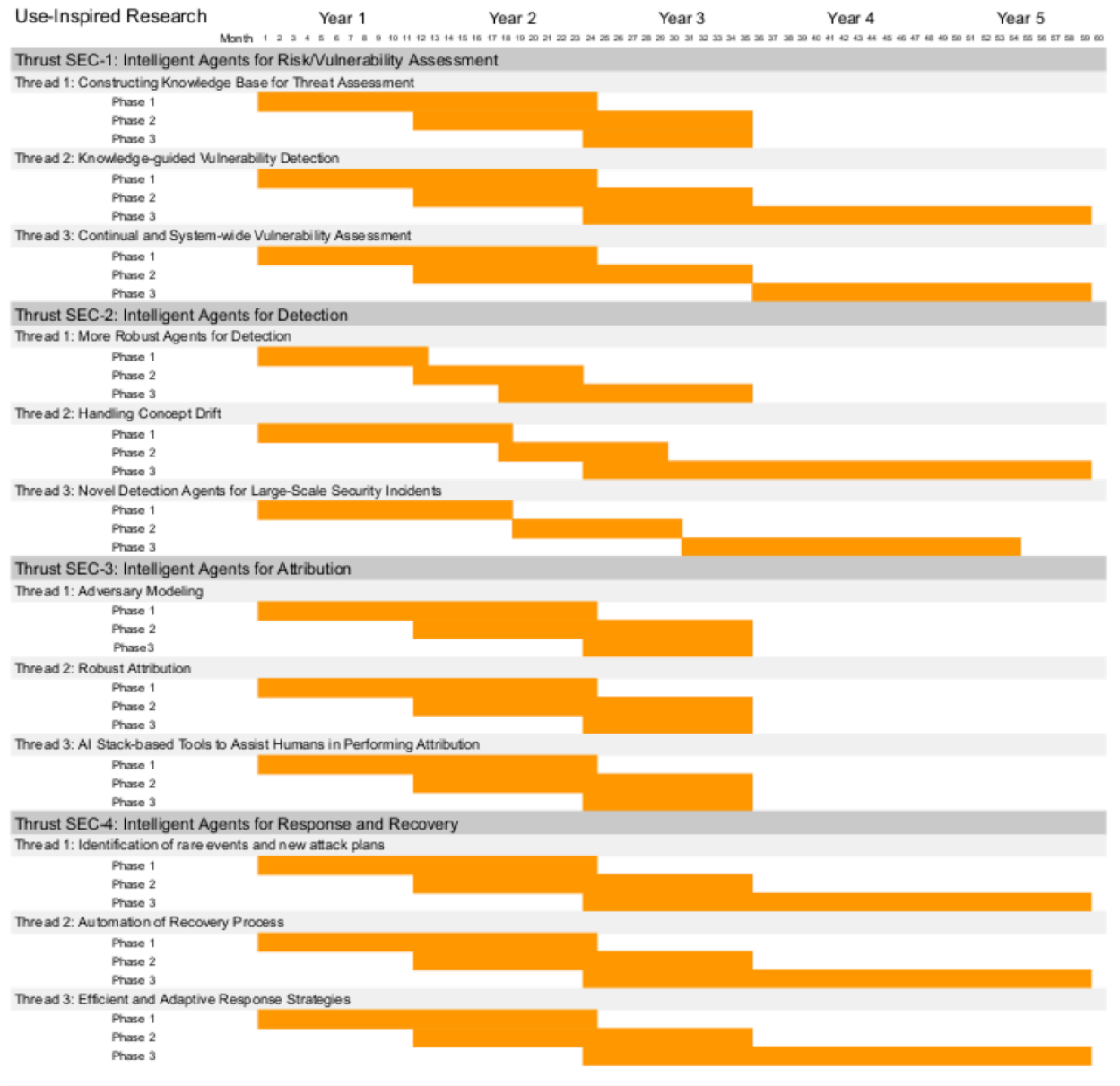


Figure 2: Timeline of security, use-inspired research activities.

Notice that initial research activities occur during the first 24 months of ACTION. Accordingly, most of the activities currently underway during this reporting period and described in this report are on-going, and there will be additional updates during our next annual report.

This implementation strategy has also allowed researchers across the Institute to finalize or progress upon research that may have begun prior to ACTION's inception. The extended, 2-year phase also allows researchers to familiarize themselves with each other's work.

3.1.1 Foundational AI research

ACTION's first year of activity was marked by connections made among the Institute's researchers. In many cases, collaborations occurred that involved investigators in areas where they were not initially expected to contribute. For example, working with **Dr. Bo Li** brought **Dr. Radha Poovendran** into Thrust AI-1, on informing a knowledge base for the proposed AI stack. Other security researchers (**Dr. Gang Wang** and **Dr. David Wagner**, for example) made contributions to activities informing Thrust AI-2, on human-in-the-loop decision-making. Indeed, we believe it is the novel combination of top researchers in both AI and cybersecurity that is allowing us to overcome the perceived challenges within each of our research activities.

Appendix B highlights the contributions and achievements of each of our research activities. Our most successful activities revolved around jailbreaks for LLMs, and how knowledge of these attacks helps in the development of a unified knowledge base for future intelligent models (Thrust AI-1 and SEC-1). There has also been significant progress in adversarial modeling and threat assessment and detection (Thrust AI-3 and AI-4, and SEC-2). AI researchers continue their work on human-agent teaming, and modeling the influence AI agents have on human decision making.

In the past six months we have also made significant progress on our GATE infrastructure. By the end of May, we were able to demonstrate a network traversal attack, designed by researchers at Georgia Institute of Technology, against the model infrastructure of our fictional city, New Esperanza. The attack allowed us to refine the logging infrastructure and re-architect the standard VPC. A separate effort to simulate cyber-physical systems in the cloud was developed and implemented by researchers at Purdue. Researchers from several ACTION campuses contribute to GATE.

A notable achievement of the use-inspired cybersecurity thrusts is the body of research on models for misattribution (SEC-3). We note a number of findings that concern data poisoning and other malicious actions. These findings will inform the activities of SEC-1 and AI-1, as we continue to refine the abstract reasoning and detection capabilities of novel agents.

A complete list of publications associated with ACTION is available on ACTION's Google Scholar profile. The list includes work submitted to ArXiv as well as publications accepted to various conferences and journals. Summer 2024 is going to be especially rich for dissemination of ACTION activities. Our investigators have more than 35 conference papers and posters to present before September. A large number of these will be published in the proceedings of the Computer Vision and Pattern Recognition Conference (CVPR 2024), to be held in Seattle, WA, in June and at the International Conference on Machine Learning (ICML 2024), in Vienna, Austria in July.

3.1.2 Other Institute research

ACTION's research plans during the first program year were intentionally broad. As such, most of the reportable activities fit into the various research thrusts. Nonetheless, there were a couple of highlights, not directly resulting from ACTION, but that play an important role in the tone the Institute will take in the development of the AI stack – namely one of transparency and responsibility.

For instance, in an article published in *Science* [1], **Dr. Stuart Russell** discusses government regulation of advanced intelligent agents. **Dr. Russell** also provided written and oral testimony for the U.S. Senate on the subject of AI regulation, and published a paper wherein he and his co-authors propose that any organization developing a foundation model intended for public use must demonstrate a reliable detection mechanism for the content it generates, as a condition of its public release. They then demonstrate the feasibility and benefits of such a tool.

As the Institute matures, we will provide similar updates on other institute research.

3.1.3 Benefits of use-inspired research

One example of ACTION's use-inspired research reaching beyond the scope of the project is a demonstration of continuous learning [2] where **Dr. Bing Liu** and co-authors combine continuous learning with several audio encoder pre-training methods. The audio encoders, when pre-trained continually over a sequence of multi-modal tasks, namely audiovisual and audio-text, exhibit improved performance across various downstream tasks compared to their non-continual learning counterparts, due to knowledge accumulation. The audio encoders are also capable of performing cross-modal tasks of all learned modalities. In future reports, we foresee being able to report on similar impacts.

Another use-inspired impact this year concerns the DARPA AIxCC challenge [3]. Researchers at UCSB, led by **Dr. Giovanni Vigna and Dr. Christopher Kruegel**, have entered this competition, where the applied research for the automated detection and mitigation of attacks stems from work underway at ACTION.

Vulnerability detection and attribution are paramount in another DARPA Award (#N66001-22-2-4037, PI Tevfik Bultan) involving ACTION co-PIs **Dr. Vigna and Dr. Kruegel**. This project, which also includes researchers from Purdue University, uses AI and ML to harden software systems against emergent chains of exploits that can cripple critical infrastructure.

One result is DeepCapa, an LLM framework that works on the memory snapshot of a malicious process and extracts potential malicious capabilities that the malware might execute. An important feature of DeepCapa is its ability to explain its decisions by pointing at the code that contributes significantly to the detection of the capability. When compared to the baselines, a framework like DeepCapa is not only better at malware characterization tasks but also shows improved robustness.

Our use-inspired thrusts are always allowing for connections between ACTION investigators and partners beyond the Institute. **Dr. Elisa Bertino and collaborators at Purdue University,** together with IBM Watson, have developed a transfer learning approach, based on generative adversarial networks (GAN)-domain adaptation (DA) combined with a variational autoencoder, by which labeled data from a source domain can be adapted to train a classifier in a target domain that has similarity with the source domain but only unlabeled data. Moreover, the team has also worked with Cisco research to explore their transfer learning approach for commercial security applications [4]. **Dr. Gang Wang's** collaborations with the National Center for Supercomputing Applications (NCSA) and eBay set the stage for future years of ACTION, and a more-defined GATE, where advanced prototypes will be transferred from our environment into a larger, industrial setting.

3.2 Nurture and grow the next generation of talent

Like our research activities, the ACTION Institute's efforts promoting broadened participation involved minimal deviation from what we describe in our SIP.

3.2.1. Pre-college education

Capture the Flag Exercises as a tool for K-12 Education. **Dr. Giovanni Vigna** launched the 20th anniversary International Capture the Flag (iCTF) [5] tailored toward high school students. Different from prior installations of iCTF, which tended to follow an attack/defense format, the 2023 iCTF adopted a "Jeopardy" format focused on aspects of AI, including trustworthy AI and

human-agent teaming, among other subjects. Also different from previous installations of iCTF, this year's version lasted over a period of one week, in an attempt to give high school students a chance to work on the activities in class, or after school. iCTF, like many other CTF exercises, usually takes place over a 24-48 hour period.

In October 2023, ACTION partnered with the UCSB chapter of Women in Computer Science (WiCS) to prepare and promote iCTF 2023. The UCSB WiCS chapter worked with WiCS chapters at other ACTION campuses to build a social network for dissemination of promotional materials. With ACTION's help, the team reached out to scores of high schools to promote the CTF. Ten high school teams participated. Informal feedback from the contacted schools points to a lack of familiarity with the CTF format. While hackathons are very popular, hacking challenges are not.

Our target for participation in future years is to add at least 10 high schools per year.

One of the lessons learned in setting up and running the 2023 event is that high school students require significantly more time to navigate hacking exercises. They require more time to prepare as well. To accommodate, we have already begun advertising for 2024 iCTF, which will happen in mid-November 2024.

We have also begun to package the results of the 2023 iCTF into a series of modules for use in high school classrooms. These modules, which are already described in our SIP, will not only serve as a teaching tool for computer security, but it will also be used to prepare potential high school teams for the 2024 iCTF.

High School Outreach. In addition to the training modules, another tactic within ACTION's strategy for improving pre-college education is to have graduate and undergraduate students tutor in nearby high schools. Our goal is to use cybersecurity to teach the basic principles of computer science. These plans overlap with our plans for increasing diversity in computer science. We discuss them in detail in our SIP, as well as in section 3.3.2 of this report.

We have already begun outreach in Santa Barbara High School and have plans to help teachers and students there develop a cybersecurity student club during the next reporting period. In future reporting periods, we intend to reproduce our efforts at other high schools near ACTION campuses.

Internships for high school students. Finally, **Dr. Christopher Kruegel** hosted a high school student intern from June 2023 through August 2023. Working alongside a graduate student, the intern learned about security along decentralized networks, as well as some fundamental computer science concepts such as computer hardware, cryptographic primitives, distributed systems, etc. As this was the first ACTION internship, and since the participant was a high school student, the expectations for research deliverables were minimal. However, the experience

has helped the ACTION team prepare for a larger internship program in 2024, which will include undergraduate and graduate students.

Dr. Kruegel will support the same high school student for a second summer internship in 2024. The Institute is preparing for the possibility that additional high school internships could spring from the K-12 educational plans being implemented by ACTION.

3.2.2 University, College, and Community College Education

Capture the Flag exercises as a tool for undergraduate education. We hosted a track for undergraduates in the iCTF in December 2023. iCTF allows ACTION to impact hundreds of students at a time. There were 38 undergraduate teams participating from schools around the country, with up to 10 students on each team. The winning team hailed from Purdue University.

Our goal for iCTF 2024 is to grow this number by at least ten more schools.

WiCS helped with implementation and promotion of the undergraduate iCTF. A number of graduate students from ACTION research labs, especially those students present on the Student Advisory Council, helped to develop the challenges used in the competition.

New graduate and undergraduate courses. One of our goals outlined in the SIP is to create new courses for undergraduate and graduate students, specifically on AI and computer security. This program year, we collected 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. The idea is to use the resources in place to find potential collaborations and then 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.

We will look further into the development of a Massive Open Online Course (MOOC) that teaches basic aspects of AI and computer security. Alternatively, we have explored filming classes and making them available to the public via YouTube or other video platforms. **Dr. Nick Feamster** volunteered to formally lead the course development component in Year 2.

Summer research experience for undergraduate and graduate students. We launched our inaugural summer research experience in AI+Security for undergraduate and graduate students in Summer 2024. By November, we had developed a FERPA-compliant online portal for receiving and reviewing applications. This portal tracks applicant information, including demographic information, and allows seamless review of submissions by the ACTION review committee. We

utilized connections among ACTION research faculty and the ACTION Student Advisory Council (SAC) to advertise our program. We also promoted the internship among a number of state universities and city colleges – especially those containing cybersecurity majors and those listed as National Centers of Excellence in Cybersecurity [6].

By late February, we had 113 applications for 12 openings. By late March, we had determined a short-list of participants. In April, after consulting with the graduate students and postdoctoral scholars who would be advising the ACTION projects, we selected eight undergraduate and four graduate participants.

The 12 participants will receive stipend and subsistence support during their 8-week internship. ACTION management worked with UCSB's Housing Office to secure room and board for any visiting interns.

We will describe all projects in detail in our next annual report.

Individual faculty supporting undergraduates. In Fall 2023 **Dr. Elisa Bertino and Dr. Ninghui Li** involved an undergraduate student (Yingjun Lin) in their project on robustness research (Thrust SEC-2, Research Thread 1).

Also, as part of her 2024 spring course on Data Security and Privacy, **Dr. Elisa Bertino** and the course TA have designed two projects focusing on the use of LLMs (specifically ChatGPT) for security tasks. The first project deals with generating access control commands (like GRANT and REVOKE) to implement access control policies; the second deals with detecting different types of vulnerabilities in Java code. These projects are designed to get the students acquainted with the proper use of LLMs for security tasks.

Postdoctoral Scholars.

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

3.2.3 Workforce development

The ACTION Institute's cross-institutional seminar leads to conversations about what graduate students should expect when seeking academic positions at R1 institutions. ACTION faculty intend to hold similar talks during future reporting periods.

In our SIP we discuss how the iCTF modules and MOOCs under development in future years will be tailored toward workforce training. We hope to report more on this impact in future reports.

3.2.4 Informal Learning and Education

Unrelated to ACTION's efforts to create high school programs, but still impactful for K-12 outreach, **Dr. Radha Poovendran** of the University of Washington was approached by *Science News Explores*, [7] an educational magazine and website for middle school and high school students, to run a story about red-teaming LLMs. The story [8] frames collaborative research by Fengqing Jiang, who is a member of the ACTION Student Advisory Council, along with **Dr. Poovendran**, **Dr. Bo Li,** and Dr. Bhaskar Ramasubramanian from Western Washington University in Bellingham, in a way that is approachable for young readers.

3.3 Broadening participation and diversity

The Institute's plans for broadening participation and diversity are consistent with the plans laid out in the revised SIP. 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 make computer science and cybersecurity more accessible to historically underrepresented communities. While ACTION representatives at all participating campuses will continue to promote the Institute among groups that advocate and support these communities (e.g. the Society for the Advancement of Chicanos/Hispanics and Native Americans in Science – SACNAS), our three primary strategies for broadening participation are the iCTF, in-person outreach at local high schools, and our summer internship program. We also have begun multiple explorations into collaborating with Minority Serving Institutions (MSI) on ExpandAI concept designs and proposals.

3.3.1 International Capture the Flag

iCTF is described above in Section 3.2.1 as well as in the SIP; the CTF does not require additional description here. We believe the activity helps us achieve our goals of broadening participation in that the CTF allows us to bring our research to a wide audience. Our plans to create educational modules on success in a CTF will allow us to enrich the groups and communities who typically do not have access to this sort of training.

3.3.2 High School Outreach

To help with our plans for promoting diversity at the K-12 level, this year we implemented plans for maintaining a presence at local high schools. Our efforts began in February 2024 when UCSB computer security researchers, Dr. Ilya Grishchenko and PhD Student Stijn Pletinckx, led an engaging, three-Saturday workshop through a UCSB Community outreach program called School for Scientific Thought (SST), which collaborates with UCSB's Early Academic Outreach Program (EAOP). The Saturday morning clinics provided hands-on experience with innovative tools like ChatGPT to groups of students who were bussed to UCSB from various schools, some of which were two hours away. Through interactive sessions, Grishchenko and Pletinckx equipped the students with the technical skills needed to tackle elementary security challenges, and to take their first steps into the world of cybersecurity and AI. In addition to introducing technical skills, Pletinckx and Grishchenko also highlighted the possible career paths that exist within the fields of cybersecurity and AI, and on how a focus on inclusivity and diversity is necessary to keep these two fields accessible from every background. They also emphasized the ethics discussions surrounding security, and taught the students how to be aware of, and avoid, bias in machine learning models.

The success of the SST clinic led to connections with local high school computer science instructors and further collaboration with the WiCS members who helped with iCTF 2023. We staged a computer science demonstration at Santa Barbara High School in April 2024 and now have plans to maintain a weekly presence at the school in the 2024-2025 academic year. We intend to empower students with the technical know-how to participate in CTF and other computer science activities. Our hope is that maintaining a presence on campus, and the creation of a student-led cybersecurity club will allow the computer science curriculum there to spread to students who might otherwise not be introduced to computer science.

Later in the school year, we have plans to host a clinic for teachers at other high schools so they can take the curriculum we develop and repackage it for their own communities.

ACTION's efforts toward broadening participation have won us the 2024 Faculty Outreach Grant [9], awarded by UCSB's Chancellor's Outreach Advisory Board. The $20k grant will help fund the outreach efforts of ACTION PhD student, Stijn Pletinckx, and provide stipends for high school teachers to attend the cybersecurity clinic. We are currently leveraging this award to deepen our relationship with the Santa Barbara County Education Office, with hopes of increasing our impact across neighboring counties and California. By Year 3 of ACTION, we hope to have cybersecurity clubs at high schools near each ACTION partnering campus.

3.3.3 Summer Research Experiences

Our summer research experiences help our broadening participation objectives in that we pay special attention to applicant background and opportunities. Indeed, applications give additional space for interested students to discuss hardships and barriers that have been overcome. Being on the application, responses to these questions become criteria for consideration by application reviewers. The research experience becomes more about the applicants, and how participating in ACTION research can benefit them, than about how the applicants can benefit ACTION.

We promote our program primarily through state colleges and universities and community colleges, which are all known for higher populations of underrepresented and first-generation students, and aim for 50% participation from underrepresented groups in our summer programs for undergraduate students. We received 68 undergraduate applications for Summer 2024.

3.4 Multidisciplinary integration and community building

3.4.1 Seminars

The most prominent evidence of the Institute's community building is in the various seminars hosted by ACTION. These lectures are offered as hybrid events and advertised across all campuses in the Institute, and also to sponsoring organizations. ACTION hosts two types of seminars: distinguished lectures for experts in the field; and 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. [10]

Our goal is to hold at least one distinguished lecture per academic quarter, starting in Winter 2024. We doubled the number of presentations in Spring 2024 to make up for a lack of presentations in Fall 2023. Below is a list of 2023-2024 seminars.

Distinguished Lectures.

- February 22, 2024. "Making Adversarial ML Practical." Nicholas Carlini, Google DeepMind.
- May 2, 2024. "Go Hack Yourself: More War Stories from ~50k Pentests." Snehal Antani, Horizon3.
- May 31, 2024. "The Mythical LLM-Month: Building Trustworthy Copilots and AutoPilots." Claudionor Coelho. Zscaler, Inc.

Student seminars.

- November 17, 2023. “Unforgeability in Stochastic Gradient Descent.” Teodora Baluta, National University of Singapore.
- March 14, 2024. “Automatic Malicious Code Generation Using GPT Models.” Marcus Botacin, Texas A&M University.
- March 28, 2024. “Toward Provable and Explainable AI Security.” Chong Xiang, Princeton University.

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

Finally, ACTION encourages faculty researchers to cross-pollinate through lectures at partnering campuses. **Dr. Berkay Celik** spoke to **Dr. David Evans’s** Computer Science Department at the University of Virginia on December 1, 2023. His lecture was titled, “Towards Compositional Secure Autonomy: From Perception to Control.”

3.4.2 Campus Events

The ACTION Institute sponsored the 2023 Responsible Machine Learning Summit [11] held at UCSB on October 20, 2023. Topics at the summit included Advances in Generative AI and its Applications, Potential Abuse of the new technology, Copyright and Attribution, Co-existing with AI: The Human Alignment Problem, as well as algorithmic solutions to these problems. Although the speakers hailed primarily from Computer Science and Engineering, the audience contained a number of interdisciplinary researchers, including faculty and graduate students from Philosophy, Political Science, and Communication, who are interested in the proliferation of generative AI. **Dr. Giovanni Vigna** delivered a talk during the opening remarks.

3.4.3 Other Meetings and Communication

ACTION maintains a website and a professional Slack workspace for sharing information and announcements. We also hold Institute-wide meetings for faculty on a quarterly basis. The first formal ACTION meeting took place in Santa Barbara, CA in August 2023, to kick off a retreat for drafting the SIP. Since then, meetings have been held virtually, at a time that is convenient for a critical mass of participants.

We ask that thrust leads also organize smaller meetings among their teams on a monthly or quarterly basis.

Dr. Giovanni Vigna, along with the SAC, has 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 loosely structured to give students free rein to collect papers. The ultimate goal is to prepare a systematization of knowledge report for each of the thrusts.

3.5 Multi-organizational synergies/achievements

In demonstration of multi-organizational synergies, **Dr. Xiaoming Huo** and **Dr. Bin Yu** are co-organizing a session at the Joint Statistical Meeting in August 2024 in Portland, Oregon [12]. The session title is "Statistical Methods in Cybersecurity." Speakers are **Dr. Jie Gao, Dr. Bo Li, Dr. João Hespanha, and Dr. Radha Poovendran.**

Dr. Radha Poovendran and **Dr. Bo Li** also collaborated to organize the Trojan Detection Challenge 2023 (LLM Edition), a NeurIPS 2023 Competition. The competition aimed to advance the understanding and development of methods for detecting hidden functionality in large language models (LLMs). The competition featured two main tracks: the Trojan Detection Track and the Red Teaming Track. In the former, participants were given LLMs containing hundreds of Trojans and tasked with discovering the triggers for these Trojans. In the Red Teaming Track, participants were challenged to develop automated red teaming methods that elicit specific undesirable behaviors from an LLM model fine-tuned to avoid those behaviors. These activities mirror Thrust AI-3 and AI-4.

The 2023 iCTF involved collaboration among ACTION institutions. During the month preceding the security exercise, the six participating ACTION institutions familiar with security exercises generated 26 challenges (11 high school and 15 undergraduate). Each institution worked with a lead graduate student from UCSB to create and deploy their challenge. The iCTF was played by 38 undergraduate and 10 high school teams from across the country. The winning undergraduate team hailed from Purdue University. The winning high school team was from Çorlu Arif Nihat Asya Anadolu Lisesi, in Turkey. The second place high school was Mountain View High School in Stafford County, Virginia.

Finally, although not "multi-organizational," we feel it is a noteworthy achievement that Co-PI **Dr. Christopher Kruegel** was elected to the 2023 Class of ACM Fellows for "contributions to systems security, malware detection, and vulnerability analysis."

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

- **Jie Gao**. "Enabling Asymptotic Truth Learning in a Network." Invited Paper for "Statistical Methods in Cyber Security." at the 2024 Joint Statistical Meetings (JSM). August 6, 2024.
- **Jie Gao.** "Dynamics and Learning on a Social Network." Invited talk at Department Colloquium at Department of Statistics. Rutgers University. September 13, 2023.
- **Jie Gao**. Invited talk at Workshop on Massive Data Models and Computational Geometry. Bonn, Germany, September 23-27, 2024.
- **Berkay Celik.** "Towards Compositional Secure Autonomy: From Perception to Control." Invited talk at the University of Virginia. December 1, 2023.
- **Stuart Russell.** Written testimony for the United States Senate AI Forum on Risk, Alignment, & Guarding Against Doomsday Scenarios. Washington, DC, December 6, 2023.
- **Jie Gao**. "Differentially Private Range Query on Shortest Paths." Invited talk at New York University Geometry Seminar. December 12, 2023.
- **Jie Gao.** "Graph Ricci Flow and Applications in Network Analysis and Learning." Invited talk at Joint Mathematics Meetings. San Francisco, CA, January 3-6, 2024.
- **Gang Wang.** "Is artificial intelligence the solution to cyber security threats?" Interview in *Financial Times* about generative AI and cyber security. January 2024.
- **Gang Wang.** "AI for Offense: What Does It Mean for Disinformation and Social Engineering?" Invited talk at AI and Security CoP: External Speaker Series. Intel, Virtual, February 2024.
- **Giovanni Vigna** provided presentations on ACTION to collaborating faculty from Politecnico di Milano and University of Twente, Netherlands. February 2024.
- **Jie Gao**. "Graph Ricci Flow and Applications in Network Analysis and Learning." Invited talk at Learning Meets Geometry, Graphs, and Networks. New Jersey Institute of Technology, February 29 - March 1, 2024.

- **Elisa Bertino.** "Applying Machine Learning to Securing Cellular Networks." Keynote at ACNS 2024: 22nd International Conference on Applied Cryptography and Network Security. Abu Dhabi, UAE, March 5-8, 2024.
- **Elisa Bertino** co-chaired the 2023 Individualized Cybersecurity Research (IMentor) Workshop and gave a keynote titled “The Persistent Problem of Applications Insecurity”
- **Berkay Celik**. "Unlocking the Potential of Generative AI for Security Applications." Invited talk at MITRE, AI-Cyber Convergence Technical Exchange Meeting (TEM). April 2024.
- **Berkay Celik**. "Towards Compositional Secure Autonomy: From Perception to Control." Invited talk at Akademiska Föreningen, Lund University Sweden for the ELLIIT focus period Symposium on Security and Fault Tolerance of Cyber-Physical Systems. April 2024.

3.6.3 Other Meetings Related to Knowledge Sharing

In March 2024, a team of 13 DARPA program specialists visited UCSB while conducting a series of site visits in Southern California. At their request, ACTION faculty and graduate students delivered a series of presentations on AI and Security, including an overview of ACTION, recent research on employing AI to detect chained vulnerabilities, research and plans at UCSB for AI for automated detection and patching of vulnerabilities, and AI for malware detection.

There are no formal exchanges of IP or technology transfer during our first year. However, as noted in the use-inspired section above, ACTION faculty researchers like **Dr. Elisa Bertino and Dr. Gang Wang** are each collaborating with companies such as 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 Media

Upon notice of the award in mid-2023, Co-PIs at ACTION partnering campuses raised awareness of the Institute through media outlets. UCSB, Purdue, and Georgia Tech [13]–[17] all ran stories in their college magazines, and Co-PIs were featured on local interviews. This media attention helped make ACTION a focal point for research questions related to AI. Persons and organizations interested in collaborating with academics in the area of AI and cybersecurity knew to reach out to ACTION.

As one example, in early 2024, faculty at UCSB began a Community of Practice (CoP) related to Artificial Intelligence. The faculty coordinators of this CoP naturally sought out ACTION's assistance with presentations during the preliminary meetings, including one presentation related to our efforts regarding the introduction of computer security concepts to impact pre-college education.

3.7.2 New grants

Dr. David Wagner and **Dr. Christopher Kruegel** submitted a proposal on AI for Cybersecurity to the UC Noyce Initiative [18]. The proposal was accepted for funding and reflects the collaborative research of ACTION faculty. A third Co-PI on this project is Hao Chen from the University of California, Davis.

3.7.3 Other collaborative efforts

For their roles in ACTION, Dr. Giovanni Vigna and Tim Robinson were invited to the second annual Summit of Centers for Digital Trust (C4DT), August 28-30, in Glion Switzerland. The summit is hosted by the Swiss Federal Institute of Technology (EPFL) and will promote the gathering of a global network of research centers to foster connections and strengthen their impact on industry and government agendas. One of the working groups at the summit will highlight capture the flag (CTF), which is especially relevant to ACTION's activities.

ACTION faculty researchers continue to collaborate with one another on the various research activities outlined in this report and the Institute's SIP.

4.0 Evaluation and Assessment

ACTION's external evaluator, Dr. Tarek Azzam of UC Santa Barbara, helped design the applications for iCTF and the summer research experience, collected survey data from participants, and prepared evaluative reports the Institute uses to improve its programs. The evaluation covers four areas: outreach and mentorship, academic preparation, research preparation, and professional development and networking.

5.0 Infrastructure and Management

5.1. Configuration and Leadership

5.1.1 Organizational Chart

During its first year, the ACTION Institute maintained the organizational structure described in the SIP. We have revisualized our organizational chart below.

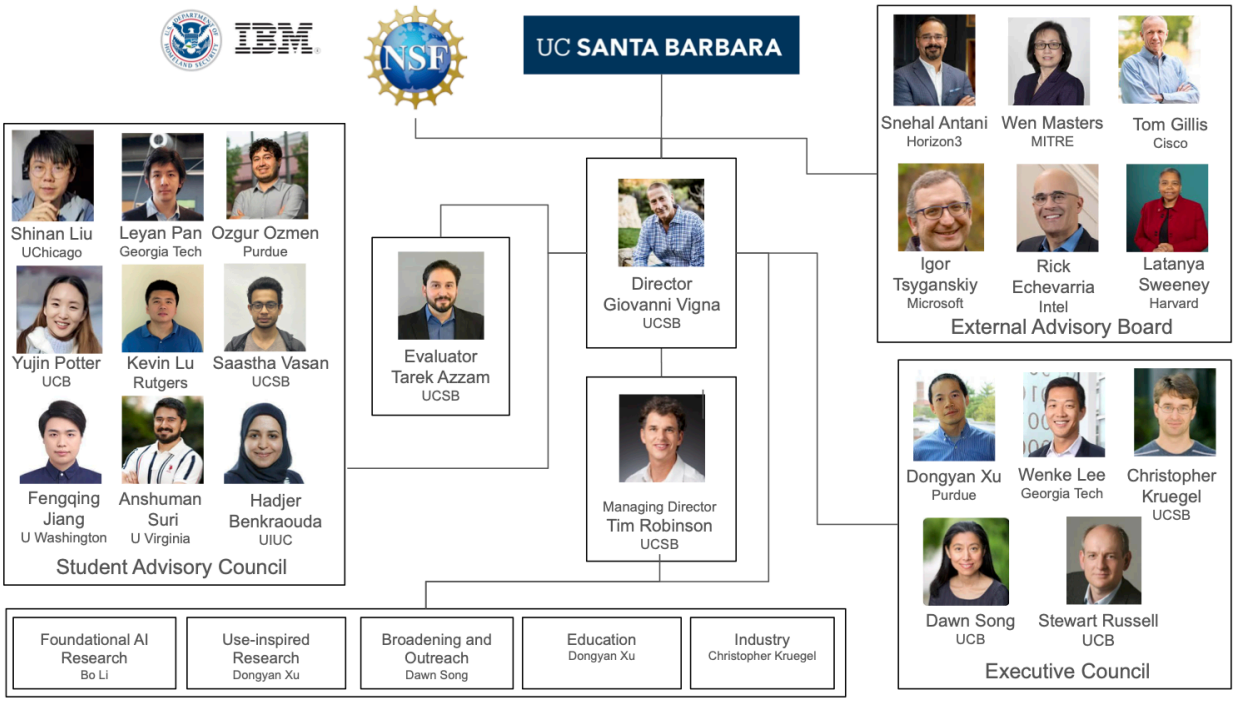


Figure 3. ACTION Institute Management Organizational Chart

Brief organizational updates follow:

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

Student Advisory Council. The SAC has maintained a monthly meeting time and has elected a Council Head, **Shinan Liu** from University of Chicago. The Council prepared a charter in November 2023 and ratified the document in early January 2024.

SAC members have been very active in Institute activities. They helped with the initial round of summer internship application review in February and March 2024. They helped select speakers for the quarterly ACTION seminars, inviting them when appropriate. **Council Head Liu** often initiates the hosting of seminars and runs council meetings. The SAC helped with promotion of iCTF 2023 and those members with CTF experience participated in challenge design. The SAC has also been heavily involved in the SoK meetings launched in late March 2024.

External Advisory Board. The EAB met in April 2024 and provided commendations and recommendations.

5.1.2 Ethics Training

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

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 SIP, have not changed. Both advisory boards have agreed upon charters to define their operations. Given that ACTION is just now ending its first year of activity, the advisory boards are relatively new.

5.2.1.2 Project management

ACTION's SIP identifies the projects needed in the Institute's first two years, which form the foundation for later work, as well as which broadening participation activities should come first. This year, faculty pursued projects that lay the groundwork for the early foundational and use-inspired research phases.

The immediate research agenda involves identifying the current state of the art for an AI stack and deploying agents in the GATE infrastructure. The stack is based on the eight research thrusts, and its team leads are the thrust leads named in the SIP. ACTION researchers are close to deploying agents in GATE. To broaden participation in developing the AI stack, researchers have also begun collecting literature for a series of systematization of knowledge (SoK) papers. Dr. Giovanni Vigna leads weekly SoK meetings.

The thrust leads and Executive Committee members were selected when the ACTION proposal was drafted. The one change is Dr. Bo Li, who during Year 1 succeeded Dr. Stuart Russell as lead of Thrust AI-1. Teams within each thrust form around the interests and collaborations of individual faculty, and ACTION encourages these organic collaborations by creating opportunities for researchers to come together. Faculty identify related projects and complementary research in their own departments and bring them to the Executive Committee, with support from ACTION management.

5.2.1.3 The role of mentors

The standard relationship between faculty advisor and student/postdoc remains in place in the ACTION Institute. As he is leading the SoK meetings and maintaining a presence during the SAC meetings, **Dr. Giovanni Vigna** has become something of a mentor as well.

ACTION maintains an open, safe space for dialogue. Mentees are encouraged to ask questions and seek advice.

Postdocs have a more-specific mentoring plan that has been laid out in the SIP. When the postdoctoral fellowship program gets underway in Year 2, we will have more to report on postdoctoral mentoring.

5.2.2 Conflicts of Interest

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

5.3 Resources and Organizational Equipment

Faculty across the institute conduct ACTION activities in conjunction with current research.

References

[1] M. K. Cohen, N. Kolt, Y. Bengio, G. K. Hadfield, and S. Russell, "Regulating advanced artificial agents," *Science*, vol. 384, no. 6691, pp. 36–38, 2024.

[2] G. Kim, H. H. Wu, L. Bondi, and B. Liu, "Multi-modal continual pre-training for audio encoders," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2024, pp. 691–695.

[3] DARPA, "AI Cyber Challenge (AIxCC)." [Online]. Available: https://aicyberchallenge.com/

[4] S. Bhardwaj, A. S. Li, M. Dave, and E. Bertino, "Overcoming the lack of labeled data: Training malware detection models using adversarial domain adaptation," *Comput. Secur.*, Art. no. 103769, 2024.

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

[6] National Security Agency, "National Centers of Academic Excellence in Cybersecurity." [Online]. Available: https://www.nsa.gov/Academics/Centers-of-Academic-Excellence/

[7] *Science News Explores*, "About us." [Online]. Available: https://www.snexplores.org/about-us

[8] "'Jailbreaks' bring out the evil side of chatbots," *Science News Explores*. [Online]. Available: https://www.snexplores.org/article/chatbot-jailbreaks-bad-ai

[9] UC Santa Barbara, "Faculty outreach grants." [Online]. Available: https://oep.ucsb.edu/initiatives/faculty-outreach-grants

[10] ACTION Institute, "ACTION Institute YouTube channel." [Online]. Available: https://www.youtube.com/@ACTIONInstitute

[11] UC Santa Barbara, "Responsible Machine Learning Summit 2023," Oct. 20, 2023. [Online]. Available: https://ml.ucsb.edu/events/summit/responsible-machine-learning-summit-2023

[12] American Statistical Association, "Joint Statistical Meetings 2024: Online program." [Online]. Available: https://ww3.aievolution.com/JSMAnnual2024/Events/pubSearchEvents

[13] University of California Office of the President, "2023 Cyber Risk Program annual report," 2023. [Online]. Available: https://security.ucop.edu/files/documents/2023-annual-report.pdf

[14] Georgia Tech School of Cybersecurity and Privacy, "A bold investment in securing tomorrow." [Online]. Available: https://scp.cc.gatech.edu/bold-investment-securing-tomorrow

[15] ACTION Institute, "Acknowledgement from Congressman Salud Carbajal." [Online]. Available: https://action.ucsb.edu/news/acknowledgement-congressman-salud-carbajal

[16] Purdue University, "NSF funds institute to research AI-powered cybersecurity," May 4, 2023. [Online]. Available: https://www.purdue.edu/newsroom/releases/2023/Q2/nsf-funds-institute-to-research-ai-powered-cybersecurity.html

[17] KEYT, "Governor Gavin Newsom signs executive order to prepare California for progress of artificial intelligence," Sep. 6, 2023. [Online]. Available: https://keyt.com/news/santa-barbara-s-county/2023/09/06/governor-gavin-newsom-signs-executive-order-to-prepare-california-for-progress-of-artificial-intelligence/

[18] UC Noyce Initiative. [Online]. Available: https://ucnoyce.org/

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

AIxCC – DARPA's AI Cyber Challenge

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

CTF – Capture the Flag (security exercise)

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

iCTF – International Capture the Flag

JSM – Joint Statistical Meeting (Institute for Mathematical Statistics)

LLM – Large Language Model

ML – Machine learning

MOOC – Massive Open Online Course

MSI – Minority Serving Institution

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

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

S&IP or SIP – Strategic and Implementation Plan

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).

TTP – Tactics, Techniques, and Procedures

UC – University of California

UCSB – University of California, Santa Barbara

VPC – Virtual Private Cloud

WiCS – Women in Computer Science

WiCyS – Women in Cybersecurity

Appendix B: Project Updates

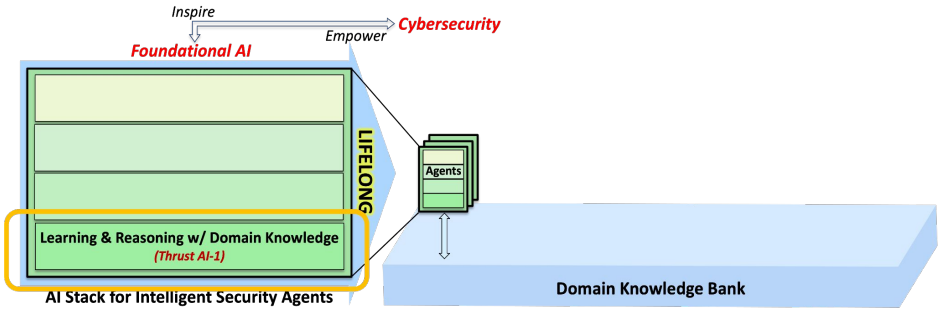
Summary slides follow.

Thrust AI-1: Learning and Reasoning with Domain Knowledge

Research Thread 1: Advanced knowledge representation learning for logical reasoning

Project Overview

- Design and optimize knowledge framework to model diverse domain knowledge.
- Identify suspicious, inconsistent predictions.
- Guide resilient decision-making.
- Phase 1 (Jul 23 - Jun 25) goal:
  - Leverage previous work to tackle problem of knowledge representation.



Overview of the ACTION Institute's integrated AI and cybersecurity research.

Challenge

- Current AI techniques cannot integrate domain knowledge and logic reasoning for prediction.

Solution

- Novel collection of ACTION researchers will expand and improve on state of the art through collaboration.

Achievements this period

- Toyer, S., Watkins, O., Mendes, E. A., Svegliato, J., Bailey, L., Wang, T., ... & Russell, S. (2024) Tensor Trust: Interpretable Prompt Injection Attacks from an Online Game.
- Jiang, F., Xu, Z., Niu, L., Xiang, Z., Ramasubramanian, B., Li, B., & Poovendran, R. (2024) ArtPrompt: ASCII Art-based Jailbreak Attacks against Aligned LLMs.
- Xu, Z., Jiang, F., Niu, L., Jia, J., Lin, B. Y., & Poovendran, R. (2024). SafeDecoding: Defending against Jailbreak Attacks via Safety-Aware Decoding.
- Xiang, Z., Jiang, F., Xiong, Z., Ramasubramanian, B., Poovendran, R., & Li, B. (2024). Badchain: Backdoor chain-of-thought prompting for large language models.

Bo Li[1,2], Stuart Russell[1,3], Pieter Abbeel[1,3], Radha Poovendran[1,4], Bin Yu[1,3]

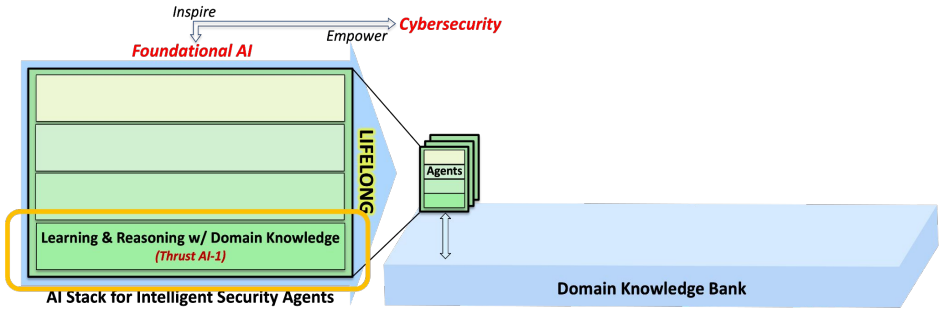
[1] Faculty, [2] University of Chicago, Computer Science, [3] UC Berkeley, Computer Science, [4] University of Washington, Electrical and Computer Engineering

Thrust AI-1: Learning and Reasoning with Domain Knowledge

Research Thread 2: Knowledge discovery through human-in-the-loop and multi-agent systems

Project Overview

- Automatically discover domain knowledge in cybersecurity tasks continually.
- Base knowledge on human feedback and interaction of agents in multi-agent system.
- Phase 1 (Jul 23 - Jun 25) goal:
  - Use models to augment AI cyber-defense.
  - Interpret human decision-making in teams and map feedback to formal knowledge rep.



Overview of the ACTION Institute's integrated AI and cybersecurity research.

Challenge

- Current Human-AI techniques cannot integrate domain knowledge and logic reasoning for prediction.

Solution

- Novel collection of ACTION researchers will expand and improve on state of the art through collaboration.
- Application of *Prospect Theory*

Achievements this period

- Prior work shows discerned behavioral shifts as individuals transition from solitary decision-making to group settings.
- Discovery of potential attacks from AI, aimed at altering risky decisions.
- Applying Prospect Theory and ML models to explore group decision-making dynamics and highlight vulnerabilities.

Bo Li[1,2], Ambuj Singh[1,3], Ming Yin[1,4]

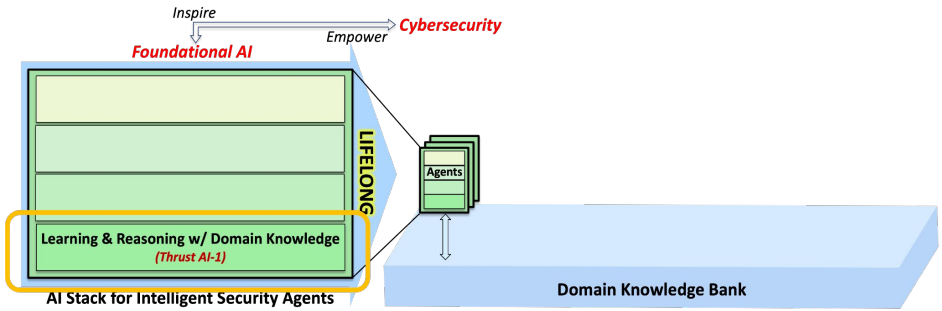
[1] Faculty, [2] University of Chicago, Computer Science, [3] UC Santa Barbara, Computer Science, [4] Purdue University, Computer Science

Thrust AI-1: Learning and Reasoning with Domain Knowledge

Research Thread 3: Veridical secure and private learning with end-to-end quality control

Project Overview

- New data/knowledge requires continuous validation.
- How to provide principled training-validation frameworks to ensure trustworthiness?
- Phase 1 (Jul 23 - Jun 25) goal:
  - Adapt PCS framework for cybersecurity data science and discovery of new knowledge through hypothesis testing.



Overview of the ACTION Institute's integrated AI and cybersecurity research.

Challenge

- Current learning approaches lack end-to-end quality control for stability and explainability of trained agents.

Solution

- Application of Yu's PCS framework for veridical data science to cybersecurity.

Achievements this period

- Kim, G., Xiao, C., Konishi, T., & Liu, B. (2023, July). Learnability and algorithm for continual learning. In International Conference on Machine Learning (pp. 16877-16896). PMLR.
- Konishi, T., Kurokawa, M., Ono, C., Ke, Z., Kim, G., & Liu, B. (2023, July). Parameter-level soft-masking for continual learning. In International Conference on Machine Learning (pp. 17492-17505). PMLR.
- Kang, M., Gürel, N. M., Li, L., & Li, B. (2024, May). COLEP: Certifiably Robust Learning-Reasoning Conformal Prediction via Probabilistic Circuits. In Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024), Vienna, Austria, May 6 - 11, 2024

Bo Li[1,2], Bin Yu[1,3], Bing Liu[1,4], Dawn Song[1,3]

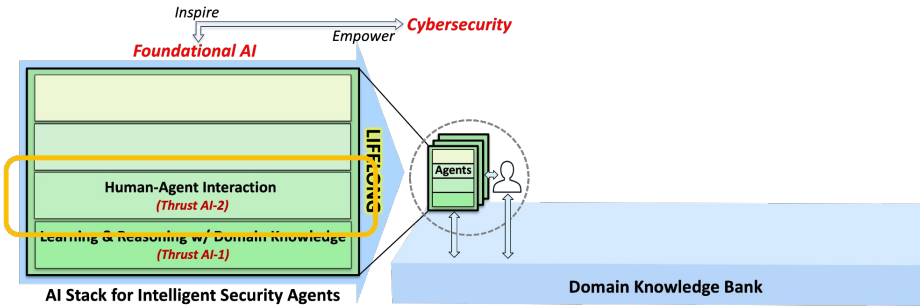
[1] Faculty, [2] University of Chicago, Computer Science, [3] UC Berkeley, Computer Science, [4] University of Illinois, Chicago, Computer Science

Thrust AI-2: Human-Agent Interaction

Research Thread 1: Design adaptive and behavior-aware AI to support human decision-making

Project Overview

- Model human behavior in human-agent teaming.
- Promote human trust in agents during decision making.
- Phase 1 (Jan 24 - Dec 26) goal:
  - Develop framework for designing adaptive AI assistance to human decision makers
  - Design adaptive algorithms to dynamically control interventions that promote trust and reliance.



Overview of the ACTION Institute's integrated AI and cybersecurity research.

Challenge

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

Solution

- Novel collection of ACTION researchers will expand and improve on state of the art through collaboration.

Achievements this period

- Piet, J., Alrashed, M., Sitawarin, C., Chen, S., Wei, Z., Sun, E., ... & Wagner, D. (2024, September). Jatmo: Prompt Injection Defense by Task-Specific Finetuning. To appear in ESORICS 2024, Poland.
- Tsai, Y. L., Hsu, C. Y., Xie, C., Lin, C. H., Chen, J. Y., Li, B., ... & Huang, C. Y. (2024, May). Ring-A-Bell! How Reliable are Concept Removal Methods for Diffusion Models? In Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024), Vienna, Austria, May 6 - 11, 2024.

Ming Yin[1,2], David Wagner [1,3], Bo Li [1,4]

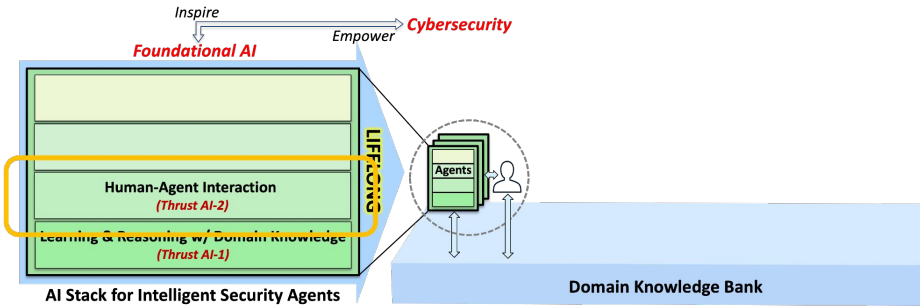
[1] Faculty, [2] Purdue University, Computer Science, [3] University of California, Berkeley, Computer Science, [4] University of Chicago, Computer Science

Thrust AI-2: Human-Agent Interaction

Research Thread 2: Efficient solicitation of human knowledge for augmenting AI

Project Overview

- Modeling human behavior in agent-in-the-loop decision-making.
- Extract domain knowledge and use it to refine AI responses to humans.
- Phase 1 (Jul 23 - Dec 25):
  - Design and refine conversational interfaces using LLM models.



Overview of the ACTION Institute's integrated AI and cybersecurity research.

Challenge

- Lack of principled ways to solicit high-level human intelligence in a scalable, reliable manner to agents.

Solution

- Novel collection of ACTION researchers will expand and improve on state of the art through collaboration.

Achievements this period

- *Lu, Z., Wang, D., & Yin, M. (2024). Does more advice help? the effects of second opinions in AI-assisted decision making.*
- Chiang, C. W., Lu, Z., Li, Z., & Yin, M. (2024, March). Enhancing AI-Assisted Group Decision Making through LLM-Powered Devil's Advocate.
- *Yuan, Y., Hao, Q., Apruzzese, G., Conti, M., & Wang, G. (2024). "Are Adversarial Phishing Webpages a Threat in Reality?" Understanding the Users' Perception of Adversarial Webpages.*
- Mink, J., Wei, M., Munyendo, C. W., Hugenberg, K., Kohno, T., Redmiles, E. M., & Wang, G. (2024, May). It's Trying Too Hard To Look Real: Deepfake Moderation Mistakes and Identity-Based Bias.

Ming Yin[1,2], Gang Wang [1,3]

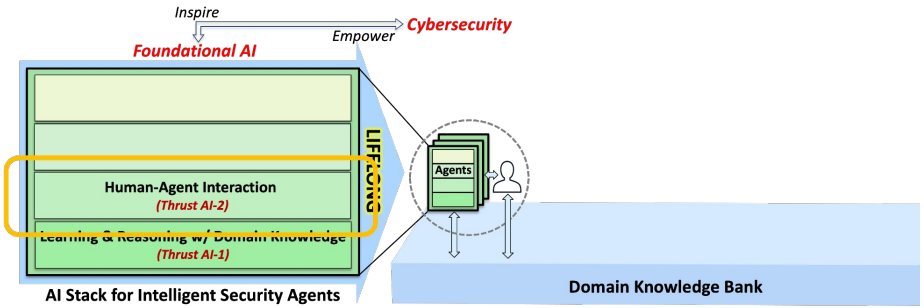
[1] Faculty, [2] Purdue University, Computer Science, [3] University of Illinois, Urbana-Champaign, Computer Science

Thrust AI-2: Human-Agent Interaction

Research Thread 3: Effective human-agent teaming

Project Overview

- Expand LLM interface framework to enable teams of humans and AI agents to communicate
- Phase 1 (Jul 25 - Jun 27):
  - Build LLM-based interfaces to enable a group of human and AI agents to communicate and negotiate the division of labor among themselves.
  - Algorithms recommend optimal division of labor.



Overview of the ACTION Institute's integrated AI and cybersecurity research.

Challenge

- Lack of ways to optimize task allocation among a team of humans and agents.

Solution

- Novel collection of ACTION researchers will expand and improve on state of the art through collaboration.

Future plans

- Our first goals within this research thread are scheduled to begin in Year 3 of ACTION.
- Interfaces will support different hierarchical structures among agents:
  - Humans and AI agents are equal teammates, each in charge of different kinds of tasks,
  - Humans serve as the "leaders" and delegate some tasks to AI agents (e.g., humans make decisions on what analyses are needed on what data and AI agents implement the analyses).

Ming Yin[1,2], Bing Liu[1,3], Ambuj Singh[1,4], Ninghui Li [1,2], Radha Poovendran [1,5], Mary Ann Hoppa [1,6], David Wagner [1,7], Bo Li [1,8], Gang Wang [1,9]

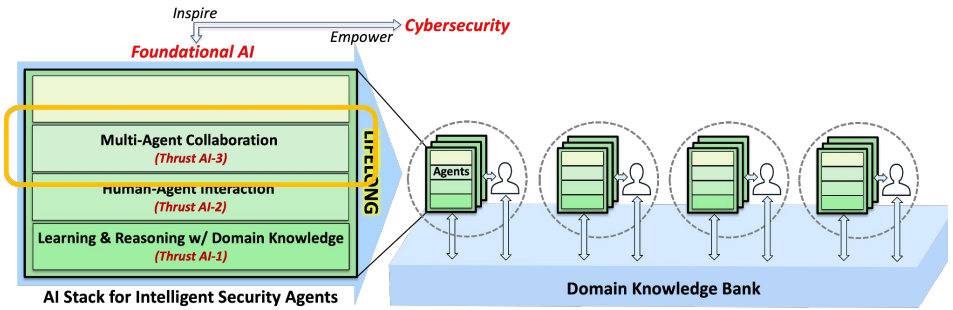
[1] Faculty, [2] Purdue, CS, [3] U Illinois, Chicago, CS, [4] UC Santa Barbara, CS, [5] U Washington, ECE, [6] Norfolk State, CS, [7] UC Berkeley, CS, [8] U Chicago, CS, [9] U Illinois, Urbana-Champaign, CS

Thrust AI-3: Multi-Agent Collaboration

Research Thread 1: Modeling of multi-agent networks

Project Overview

- Consider a distributed network of agents performing security functions.
- Co-evolving, dynamic model.
- Phase 1 (Jul 23 - Jun 25):
  - Study network-based models of multi-agents.
  - Design and analyze techniques for information exchange among agents.



Overview of the ACTION Institute's integrated AI and cybersecurity research.

Challenge

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

Solution

- Incorporate network models inspired by contagion, viral marketing, etc.

Current work and future plans

- Kosan, M., S. Verma, B. Armgaan, K. Pahwa, A. Singh, S. Medya, and S. Ranu, "GNNX-BENCH: Unravelling the Utility of Perturbation-based GNN Explainers through In-depth Benchmarking", The Twelfth International Conference on Learning Representations (ICLR), 2024.
- This thread will receive more attention during the next reporting period.

Jie Gao [1,2], Xiaoming Huo [1,3], Pieter Abbeel[1,4], Ambuj Singh[1,5], Radha Poovendran[1,6], Yung-Hsiang Lu[1,7], João Hespanha[1,8]

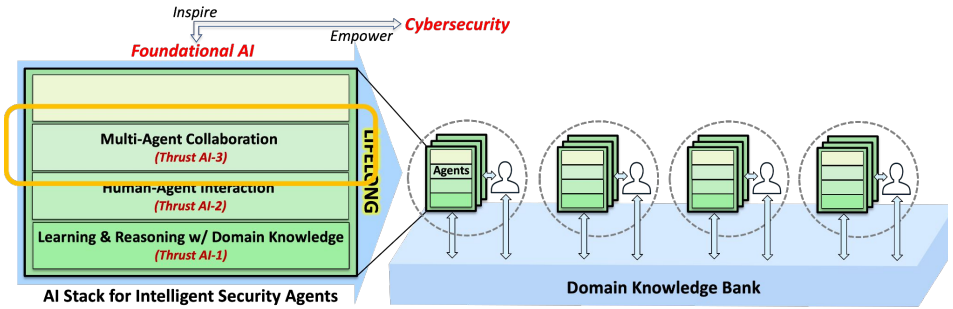
[1] Faculty, [2] Rutgers University, Computer Science, [3] Georgia Institute of Technology, Data Science and Statistics, [4] UC Berkeley, Electrical Engineering and Computer Science, [5] UC Santa Barbara, Computer Science, [6] University of Washington, Electrical and Computer Engineering, [7] Purdue University, Electrical and Computer Engineering, [8] UC Santa Barbara, Electrical and Computer Engineering

Thrust AI-3: Multi-Agent Collaboration

Research Thread 2: Resilient multi-agent reasoning

Project Overview

- Multiple agents look at different kinds of data. What are the techniques for resiliency in reasoning?
- Analysis of concept drift, optimality, federated learning.
- Designing resilient multi-agent systems that can operate despite attacks.
- Phase 1 (Jul 23 - Feb 25):
  - Create models to represent limitations of multi-agent reasoning.



Overview of the ACTION Institute's integrated AI and cybersecurity research.

Challenge

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

Solution

- Novel collection of ACTION researchers will expand and improve on state of the art through collaboration.

Achievements this period

- Jia, J., Yuan, Z., Sahabandu, D., Niu, L., Rajabi, A., Ramasubramanian, B., ... & Poovendran, R. (2024). Fedgame: A game-theoretic defense against backdoor attacks in federated learning.
- Xu, Z., Jiang, F., Niu, L., Jia, J., & Poovendran, R. (2024). Brave: Byzantine-Resilient and Privacy-Preserving Peer-to-Peer Federated Learning.
- Kang, M., Gürel, N. M., Li, L., & Li, B. (2024). COLEP: Certifiably Robust Learning-Reasoning Conformal Prediction via Probabilistic Circuits.

Jie Gao [1,2], Radha Poovendran[1,4], Yung-Hsiang Lu[1,5], João Hespanha[1,6], Wenke Lee[1,3], Xiaoming Huo [1,7]

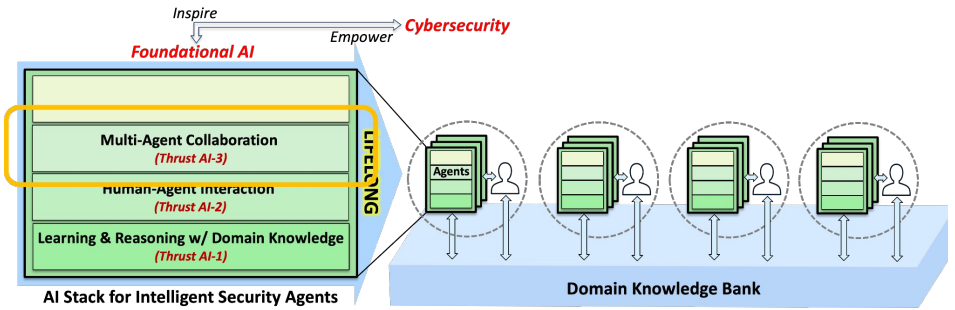
[1] Faculty, [2] Rutgers University, Computer Science, [3] Georgia Institute of Technology, Computer Science, [4] University of Washington, Electrical and Computer Engineering, [5] Purdue University, Electrical and Computer Engineering, [6] UC Santa Barbara, Electrical and Computer Engineering, [7] Georgia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

Thrust AI-3: Multi-Agent Collaboration

Research Thread 3: Robust and adversary-aware reinforcement learning

Project Overview

- Design of reinforcement algorithms for security domain using adversarial games.
- Requires formal approaches to learn attackers' intent and capabilities.
- Finite-state machines and/or discrete-event systems.
- Phase 1 (Jan 25 - Dec 27) goals:
  - Design reinforcement learning algorithms for security



Overview of the ACTION Institute's integrated AI and cybersecurity research.

Challenge

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

Solution

- Application of multi-agent reinforcement learning.

Preparatory Work

- Bhowmick, A., M. Kosan, Z. Huang, A. Singh, and S. Medya, "DGCLUSTER: A Neural Framework for Attributed Graph Clustering via Modularity Maximization", The 38th Annual AAAI Conference on Artificial Intelligence (AAAI'24), 2024.

Future Plans

- This thread will receive more focus starting in January 2025.

Jie Gao [1,2], Xiaoming Huo [1,3], Pieter Abbeel[1,4], Ambuj Singh[1,5], Radha Poovendran[1,6], Yung-Hsiang Lu[1,7], João Hespanha[1,8]

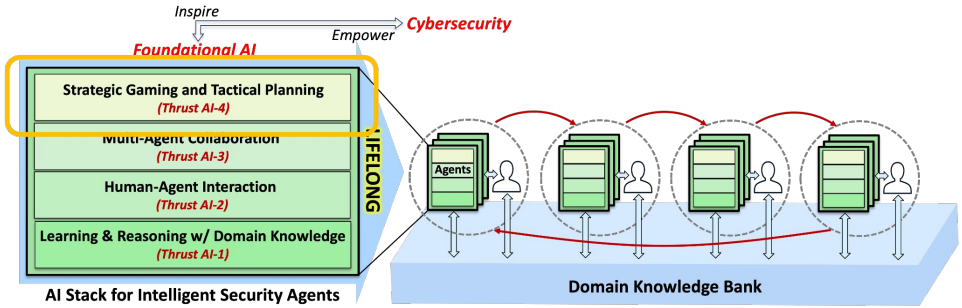
[1] Faculty, [2] Rutgers University, Computer Science, [3] Georgia Institute of Technology, Data Science and Statistics, [4] UC Berkeley, Electrical Engineering and Computer Science, [5] UC Santa Barbara, Computer Science, [6] University of Washington, Electrical and Computer Engineering, [7] Purdue University, Electrical and Computer Engineering, [8] UC Santa Barbara, Electrical and Computer Engineering

Thrust AI-4: Strategic Gaming and Tactical Planning

Research Thread 1: Integration of KG and IFG for enhancing cyber AI games

Project Overview

- Need for methods to effectively capture interactions between AI-enabled adversaries and AI-driven systems/mitigations.
- Improve detection quality by integrating additional information about adversaries and attacks.
- Current game-theoretical approaches mainly focus on defending against anticipated attacks, with little reasoning about the dynamic nature of the attackers' behavior.
- Phase 1 (Jul 23 - Jun 25) goal:
  - Develop graph structure integrating Knowledge Graphs (KG) and Information Flow Graphs (IFG) from Thrust AI-1.



Overview of the ACTION Institute's integrated AI and cybersecurity research.

Challenge

- Main solution for classical game theory is based on equilibrium.
- Cyber interactions are dynamic, often preventing convergence.

Solution

- Dynamic Information Flow Tracking (DIFT) for modeling games.

Achievements this period

- *Jiang, F., Xu, Z., Niu, L., Wang, B., Jia, J., Li, B., & Poovendran, R. (2023). Identifying and Mitigating Vulnerabilities in LLM-Integrated Applications. NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following, New Orleans, LA, Dec 10-16, 2023*

Radha Poovendran [1,2], João Hespanha[1,3], Bo Li [1,4], Giovanni Vigna[1,5], Wenke Lee[1,6]

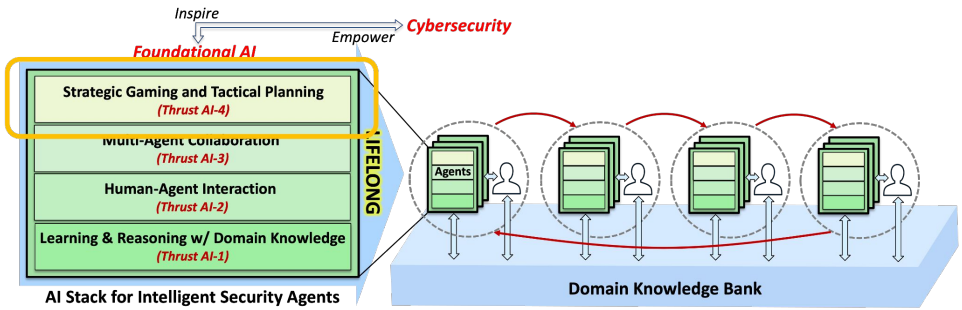
[1] Faculty, [2] University of Washington, Electrical and Computer Engineering, [3] UC Santa Barbara, Electrical and Computer Engineering, [4] University of Chicago, Computer Science, [5] UC Santa Barbara, Computer Science, [6] Georgia Institute of Technology, Computer Science

Thrust AI-4: Strategic Gaming and Tactical Planning

Research Thread 2: Adversarial modeling and certification

Project Overview

- The impact of an attack can be modeled and measured in the latent space of a graph.
- Need for efficient querying of latent space to summarize most vulnerable nodes.
- Phase 1 (Jul 23 - Jun 24) goal:
  - Model interactions between players as an adversarial reinforcement learning problem.



Overview of the ACTION Institute's integrated AI and cybersecurity research.

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.

Solution

- Instead of equilibrium, develop algorithmic approaches that would provide certifiable guarantees on the best response at each stage.

Achievements this period

- Researchers are investigating reinforcement learning problem between two sets of agents on a network over a small horizon. Exploring budget costs and payoffs.
- Markov Chain Monte Carlo (MCMC) that optimizes decision making for Koopman representation. Outperforms classical optimization algorithms for such systems; especially, for complex, large-scale dynamics.
- Data-driven algorithms for adversarial games aimed at adversarial correlation analysis in the security domain.
- Forthcoming integration into AI stack.

Radha Poovendran [1,2], João Hespanha[1,3], Ambuj Singh[1,4], Jie Gao[1,5]

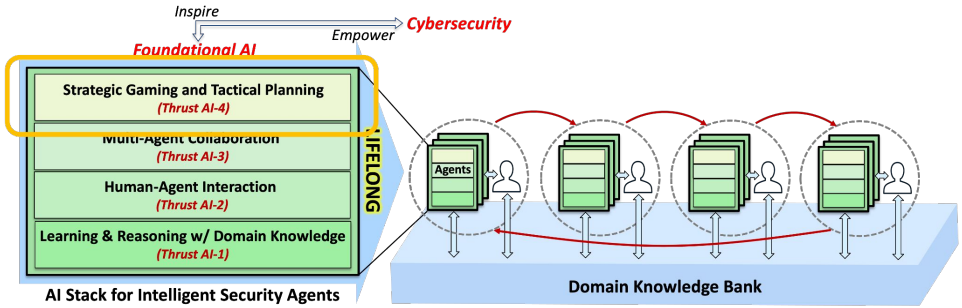
[1] Faculty, [2] University of Washington, Electrical and Computer Engineering, [3] UC Santa Barbara, Electrical and Computer Engineering, [4] UC Santa Barbara, Computer Science, [5] Rutgers University, Computer Science

Thrust AI-4: Strategic Gaming and Tactical Planning

Research Thread 3: Deception modeling approaches

Project Overview

- Cyber Games are inherently asymmetrical. Players construct opposing views that differ from actual system state.
- Hypergames are more appropriate for cyber game modeling.
- Phase 1 (Jul 24 - Jun 26) goal:
  - Develop hypergame framework to model the strategic and deceptive interactions between AI-enabled adversaries and the AI defense in Cyber AI games.
  - Identify classes of hypergames for which it will be possible to determine best response.



Overview of the ACTION Institute's integrated AI and cybersecurity research.

Challenge

- Developing precise cognitive models and maintaining deception over multiple rounds of interaction are open problems.

Solution

- Investigate computing optimal deception-based strategies in linear quadratic games with signaling.
- Develop local deception games that will provide guidance as to which agent must introduce deceptive signaling into the environment.

Future plans

- Thread requires additional preparation and survey of current literature.
- Research will begin in the next reporting period.

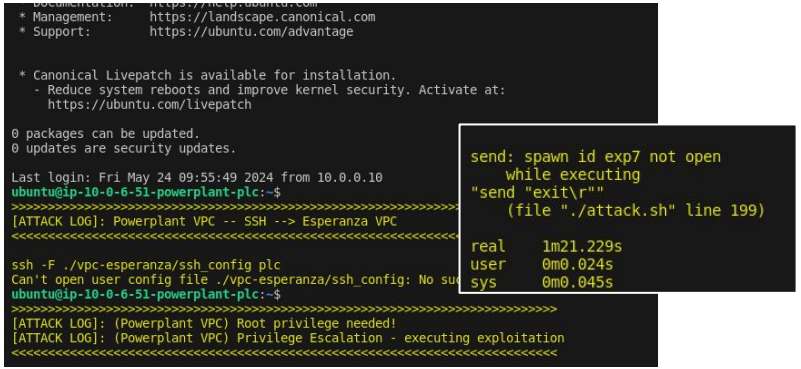
Radha Poovendran [1,2], João Hespanha[1,3], Wenke Lee[1,4]

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

Globally Accessible Testing Environment (GATE)

Project Overview and requirements

- Virtual integrated testbed for all ACTION thrusts.
- User-friendly experiment set up; always up-to-date configuration; remote access through web-based APIs.
- Flexible, extensible scenarios, based in real world.
- Phase 1 (Jul 23 - Jun 24) goals:
  - Build overarching framework.
  - Demonstrate an initial attack.



Pictured: Screenshot of attack demonstration using GATE.

Challenges

- Robust technology solutions are needed for scalability.
- Network architecture is complicated; should facilitate seamless interaction between agents and modeled humans.

Solutions

- Incremental development. Prototype-centric.
- Engage technology users early to plan for research requirements.

Achievements this reporting period

- UCSB enterprise AWS used for back-end support. Terraform development.
- Virtual model of New Esperanza facilities. Deployable PLCs. Subnetworks.
- Passive network traffic inspection and alerting.
- Initial attack and demonstration prepared.
- Truly collaborative effort across ACTION.

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

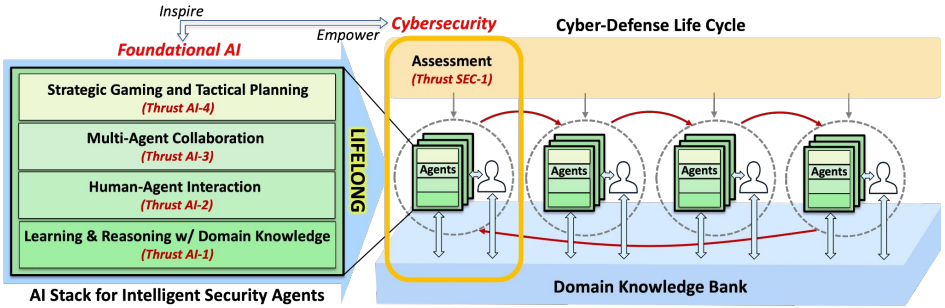
AI Institute for Agent-based Cyber Threat Intelligence and Operation

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

Research Thread 1: Constructing Knowledge Base for Threat Assessment

Project Overview

- Construct a practical, comprehensive knowledge base for threat/vulnerability assessment
- Co-developed with the foundational AI team (Thrust AI-1).
- To be hosted by GATE and replicable by the community.
- Phase 1 (Jul 23 - Jun 25) goals:
  - Survey available data sources, gather datasets, and construct the preliminary knowledge base.
  - Expand DiverseVul data set.



Overview of the ACTION Institute's integrated AI and cybersecurity research.

Challenge

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

Solution

- Leverage/integrate knowledge base from thrust AI-1.

Achievements this period

- Deep learning model to perform code-data separation for binary analysis for ICS and IoT
- Suya, F., Suri, A., Zhang, T., Hong, J., Tian, Y., & Evans, D. (2024, April). SoK: Pitfalls in evaluating black-box attacks. In SaTML 2024.
- Jiang, F., Xu, Z., Niu, L., Wang, B., Jia, J., Li, B., & Poovendran, R. (2023). Identifying and Mitigating Vulnerabilities in LLM-Integrated Applications
- Additional work on LLM jailbreaks to inform knowledge base.

Gang Wang [1,2], David Evans [1,3], Berkay Celik [1,4], David Wagner [1,5], Bo Li [1,6], Radha Poovendran [1,7], Wenke Lee [1,8]

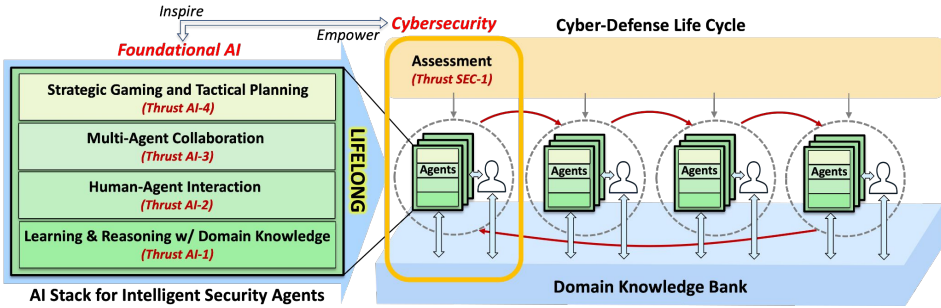
[1] Faculty, [2] University of Illinois, Urbana-Champaign, Computer Science, [3] University of Virginia, Computer Science, [4] Purdue University, Computer Science, [5] UC Berkeley, Computer Science, [6] University of Chicago, Computer Science, [7] University of Washington, Electrical and Computer Engineering, [8] Georgia Institute of Technology, Computer Science

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

Research Thread 2: Knowledge-guided vulnerability detection

Project Overview

- Knowledge-guided fuzzing to detect vulnerabilities.
- Co-developed with the foundational AI team, Thrust AI-1 and Thrust AI-2.
- Phase 1 (Jul 23 - Jun 25) goals:
  - Leverage LLM as an interface to improve accuracy, coverage, and scalability of existing CRS.
  - Use deep learning and LLMs to detect security vulnerabilities.
  - Develop new cyber-physical fuzzing methods.



Overview of the ACTION Institute's integrated AI and cybersecurity research.

Challenge

- Current techniques identify vulnerabilities without semantic/contextual system knowledge.

Solution

- Support detection with an evolving division of labor between agents and humans based on AI-2.

Achievements this period

- *Zhou, T. Y., & Huo, X. (2024). Learning ability of interpolating deep convolutional neural networks. Applied and Computational Harmonic Analysis, 68, 101582.*
- Sitawarin, C., Chang, J., Huang, D., Altoyan, W., & Wagner, D. (2024, May). Defending against transfer attacks from public models. ICLR 2024.
- Toyer, S., Watkins, O., Mendes, E. A., Svegliato, J., Bailey, L., Wang, T., ... & Russell, S. (2024, May). Tensor Trust: Interpretable Prompt Injection Attacks from an Online Game. ICLR 2024.

Gang Wang[1,2], Xiaoming Huo [1,3], David Wagner [1,4]

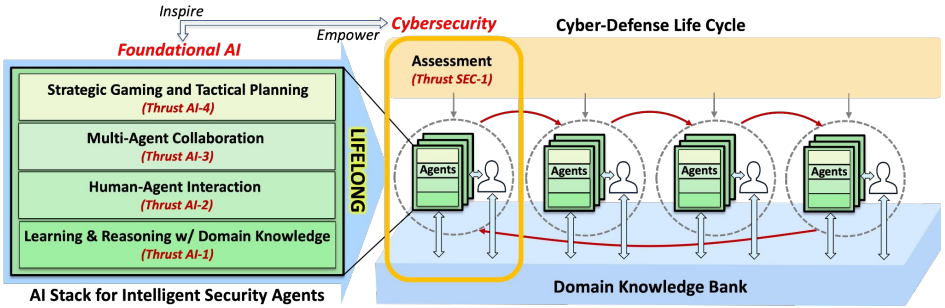
[1] Faculty, [2] University of Illinois, Urbana-Champaign, Computer Science, [3] Georgia Institute of Technology, Statistics and Data Science, [4] UC Berkeley, Computer Science

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

Research Thread 3: Continual and system-wide vulnerability assessment

Project Overview

- Focus on the challenges of dynamics in both the knowledge base and target system/software over time.
- Co-developed with the foundational AI team (Thrust AI-3).
- Phase 1 (Jul 23 - Jun 25) goals:
  - Integrate generative models and paradigms from Thrust AI-1 to improve prediction and reasoning, enhance vulnerability analysis and discovery for CPS.



Overview of the ACTION Institute's integrated AI and cybersecurity research.

Challenges

- It is difficult to revealing potential attacks automatically and proactively.
- A vulnerability in one small component may lead to other attacks.

Solution

- Formulate the attack-discovery problem as system-of-systems using cyber reasoning, in collaboration with Thrust AI-3.

Achievements this period

- Ke, Z., Liu, B., Xiong, W., Celikyilmaz, A., & Li, H. (2023, December). Sub-network Discovery and Soft-masking for Continual Learning of Mixed Tasks. *Proceedings of The 2023 Conference on Empirical Methods in Natural Language Processing (findings, EMNLP-2023), Singapore, December 6 –10, 2023*

Gang Wang [1,2], Berkay Celik [1,3]

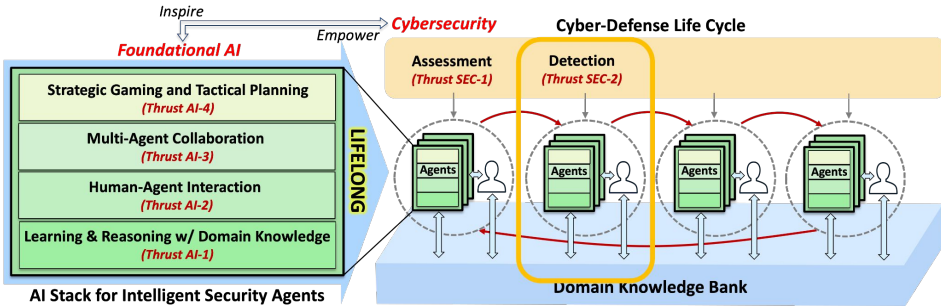
[1] Faculty, [2] University of Illinois, Urbana-Champaign, Computer Science, [3] Purdue University, Computer Science

Thrust SEC-2: Intelligent Agents for Detection

Research Thread 1: More robust agents for detection

Project Overview

- Objective: enrich domain knowledge from AI-1 and SEC-1 with high-quality info about expected behavior of system.
- Use transfer learning to train detection agents.
- Phase 1 (Jul 23 - Jun 24) goals:
  - Adversarial domain adaptation (DA) to generate domain-independent representations (DIRep).
  - Adversarially train two neural networks: discriminator and generator.



Overview of the ACTION Institute's integrated AI and cybersecurity research.

Challenge

- Current approaches degrade when facing incomplete, low-fidelity training data.
- The adversarial DA approach can suffer from the hidden data effect.

Solution

- Develop a new model combining variational auto-encoder (VAE) with the GAN-based adversarial learning approach (AI-3, AI-4).
- In new model, perform classification based on the DIRep of the data.

Achievements this period

- Development of foundational model to perform malware detection and analysis tasks.
- Release of ImageNet for Malware – malware metadata + data set.
- Novel transfer learning framework: Contrastive Adversarial Representations Distillation (CARD).
- Paper to appear at USENIX 2024: Li, J., Li, N., & Ribeiro, B. (2023). MIST: Defending Against Membership Inference Attacks Through Membership-Invariant Subspace Training.

Christopher Kruegel [1,2], Dongyan Xu [1,3], Giovanni Vigna [1,2], Elisa Bertino [1,3], Ninghui Li [1,3], Berkay Celik[1,3], Bo Li [1,4], Bin Yu [1,5]

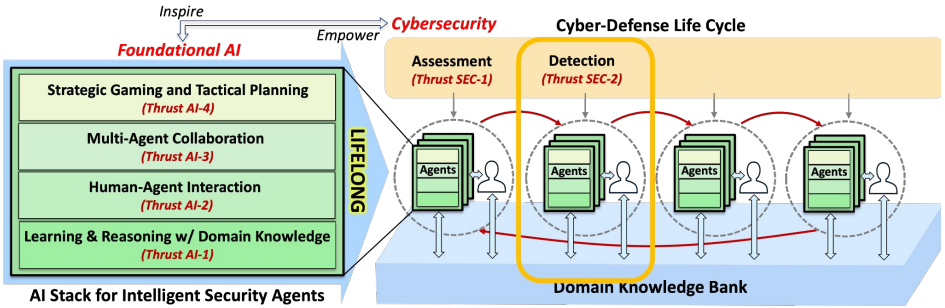
[1] Faculty, [2] UC Santa Barbara, Computer Science, [3] Purdue University, Computer Science, [4] University of Chicago, Computer Science, [5] UC Berkeley, Statistics

Thrust SEC-2: Intelligent Agents for Detection

Research Thread 2: Handling concept drift

Project Overview

- Attacker adaptation ("concept drift") is a significant challenge for security solutions.
- How to make agent detection techniques more resilient in spite of dynamic behavior and knowledge?
- Phase 1 (Jul 23 - Dec 24) goals:
  - Study techniques for detection of loss of detection accuracy.
  - Integrate a multi-dimensional analysis of underlying data into detection.
  - Incorporate discovery/feedback methods from AI-1 into AI models for detection.



Overview of the ACTION Institute's integrated AI and cybersecurity research.

Challenge

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

Solution

- Augmented feature space of ML models in Thrust AI-1 learn the relation among features during training.
- Network connections can be interpreted in the context of the users and applications that are responsible for these connections.

Achievements this period

- Framework in collaboration with NCSA to systematically detect and reason about concept drift in the network logs and inform mitigation actions.
- GAN-based domain adaptation approach for classifying drifted malware with minimal samples indicates improvement over other non-transfer learning and transfer learning schemes in actual drift malware binaries.

Christopher Kruegel [1,2], Elisa Bertino [1,3], Ninghui Li [1,3], Gang Wang[1,4]

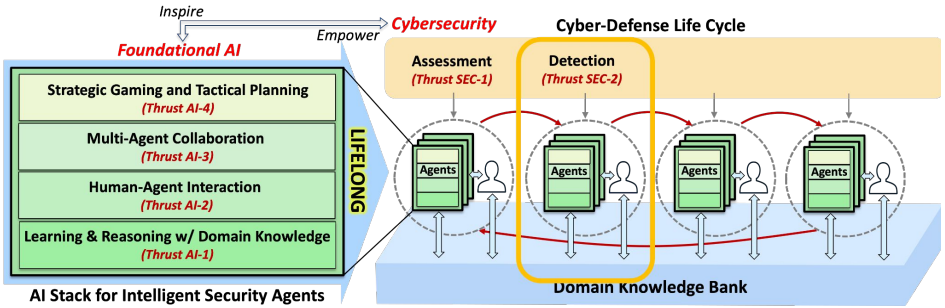
[1] Faculty, [2] UC Santa Barbara, Computer Science, [3] Purdue University, Computer Science, [4] University of Illinois, Urbana-Champaign, Computer Science

Thrust SEC-2: Intelligent Agents for Detection

Research Thread 3: Novel detection agents for large-scale security incidents

Project Overview

- Develop novel AI stack-based models to detect malware and malicious code.
- We are particularly interested in stealthy supply-chain attacks.
- Integration with GATE for testing/implementation.
- Phase 1 (Jul 23 - Dec 24) goals:
  - Develop agent that can identify potentially malicious capabilities in malware using a formal representation of the MITRE ATT&CK knowledge base.



Overview of the ACTION Institute's integrated AI and cybersecurity research.

Challenge

- Current detection methods are unreliable, resulting in low accuracy in uncertain contexts. Or, they require significant reverse engineering efforts and context-switching.

Solution

- Leverage AI-2, human-machine interaction and trust.
- Focus on alerts that "matter," inclusion of context with detection.

Achievements this period

- Collaboration with eBay to continue the investigation of methods to detect threat types that are out of distribution (OOD) of the training data. This work helps to improve the detection capability of ML models on novel threats.
- Current work to use LLMs to help security analysts ingest and search semi-structured log data.

Christopher Kruegel [1,2], Gang Wang [1,3], Elisa Bertino [1,4], Ninghui Li [1,4], David Wagner [1,5], Bo Li [1,6]

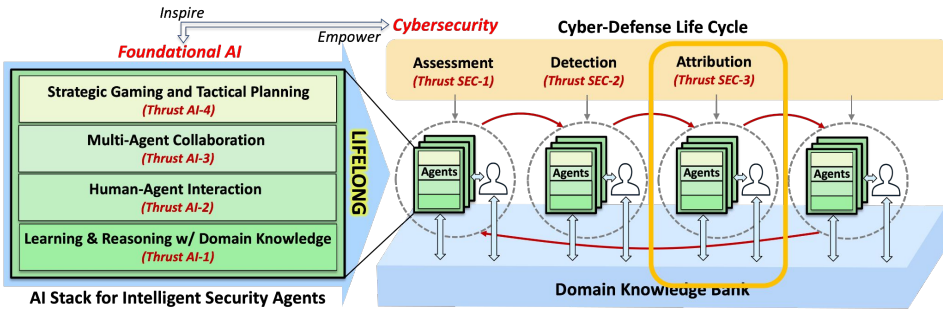
[1] Faculty, [2] UC Santa Barbara, Computer Science, [3] University of Illinois, Urbana-Champaign, Computer Science, [4] Purdue University, Computer Science, [5] UC Berkeley, Computer Science, [6] University of Chicago, Computer Science

Thrust SEC-3: Intelligent Agents for Attribution

Research Thread 1: Adversary modeling

Project Overview

- Attack attribution is necessary for prioritizing response and recovery.
- Use information from other thrusts to develop new models that capture capabilities and TTP.
- Phase 1 (Jul 23 - June 25) goals:
  - Survey different modeling approaches and integrate with various data sources.
  - Incorporate integrated graphs from AI-4 (KG+IFG).
- Phase 2 (Jul 24 - Jun 25) will focus on gleaning information from adversary mistakes to learn about attribution.



Overview of the ACTION Institute's integrated AI and cybersecurity research.

Challenge

- Current threat modeling techniques are too simple and cannot take into account collaboration among adversaries.

Solution

- AI stack (AI-1, AI-2, AI-3, AI-4) and SEC-1 present tools for developing novel agents.

Achievements this period and future plans

- Survey of modeling approaches and connection to SEC-2.
- Current development of an automated post-detection system to identify high-level, potentially malicious behaviors.
- Improvement over Forecast and CAPA.
- System can learn the rules automatically, even when the code is obfuscated.
- System relies on sequences of Windows API and library calls extracted from temporal memory snapshots taken by running malware in a sandbox environment.
- Conference submission planned for the next reporting period.

Giovanni Vigna [1,2], Christopher Kruegel [1,2]

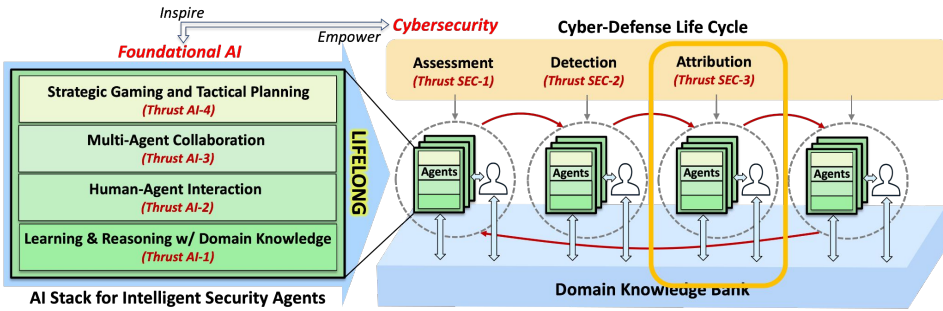
[1] Faculty, [2] UC Santa Barbara, Computer Science

Thrust SEC-3: Intelligent Agents for Attribution

Research Thread 2: Robust attribution

Project Overview

- For robust agents, we propose a “cyber-world model” that engages AI stack for prediction and can instantiate combinations of malicious actions.
- Attacks may be false flag operations.
- Phase 1 (Jul 23 - Jun 25) goals:
  - Explore causes of misattribution:
    - Lack of sufficient, high-quality data.
    - Model error.
    - Intentional deception.
  - Develop models for attribution based on AI-4.



Overview of the ACTION Institute’s integrated AI and cybersecurity research.

Challenge

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

Solution

- Integration of human analyst (AI-2) and CRS that pictures the whole system of systems (SEC-1 + AI-3) and can model/predict unforeseen consequences and potential attack sequences (AI-4).

Achievements this period

- Suya, F., Zhang, X., Tian, Y., & Evans, D. (2024). What Distributions are Robust to Indiscriminate Poisoning Attacks for Linear Learners?. Advances in neural information processing systems, 36.
- Aghakhani, H., Dai, W., Manoel, A., Fernandes, X., Kharkar, A., Kruegel, C., ... & Sim, R. (2024, May). Trojanpuzzle: Covertly poisoning code-suggestion models. IEEE Symposium on Security and Privacy, San Francisco, CA May 20-23, 2024.
- Jha, R., Hayase, J., & Oh, S. (2023). Label poisoning is all you need. Advances in Neural Information Processing Systems, 36, 71029-71052.
- Liu, X., Jain, P., Kong, W., Oh, S., & Suggala, A. (2024). Label Robust and Differentially Private Linear Regression: Computational and Statistical Efficiency. Advances in Neural Information Processing Systems, 36.

Giovanni Vigna [1,2], David Evans [1,3], Christopher Kruegel [1,2], Sewoong Oh [1,4]

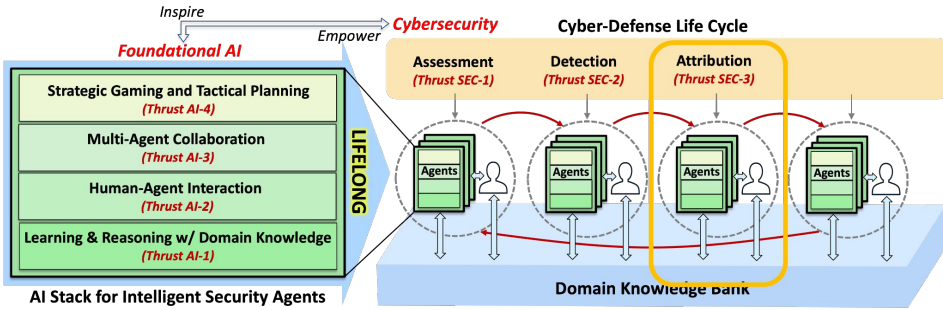
[1] Faculty, [2] UC Santa Barbara, Computer Science, [3] University of Virginia, Computer Science, [4] University of Washington, Computer Science

Thrust SEC-3: Intelligent Agents for Attribution

Research Thread 3: AI-stack tools to assist humans in performing attribution

Project Overview

- By combining LLMs with KGs, we can create intelligent agents that can provide more reliable and accurate assistance to human analysts re attribution.
- System would have a vast amount of contextual information.
- System can inform analyst about potential for evasive, misleading actions and false flag operations.
- Phase 1 (Jul 23 - Jun 25) goals:
  - Improve LLM for a more-intuitive and efficient natural language-based UI.



Overview of the ACTION Institute's integrated AI and cybersecurity research.

Challenge

- Current LLMs weaknesses: hallucination, indecisiveness, lack of domain-specific/new knowledge.

Solution

- KGs advantages: structured knowledge, accuracy, decisiveness, interpretability, and domain-specific/evolving knowledge.
- Integrated knowledge and information flow graphs (Thrusts AI-1 and AI-4) have the potential to raise alerts for further security inspection.
- Additional support/integration from AI-2 and SEC-1.

Achievements this period

- Based on ML-driven malware detection techniques, we have made improvements on large-model framework.
- Improvements utilize transfer learning techniques for downstream tasks. Reduces need for retraining for each task.
- Flexible neural network architecture can accommodate additional input during fine-tuning stage, which can further improve downstream performance.
- Future plans include exploring ways to construct a new set of fine-tuned, KG-enhanced LLMs for attribution tasks.

Giovanni Vigna [1,2], Christopher Kruegel [1,2]

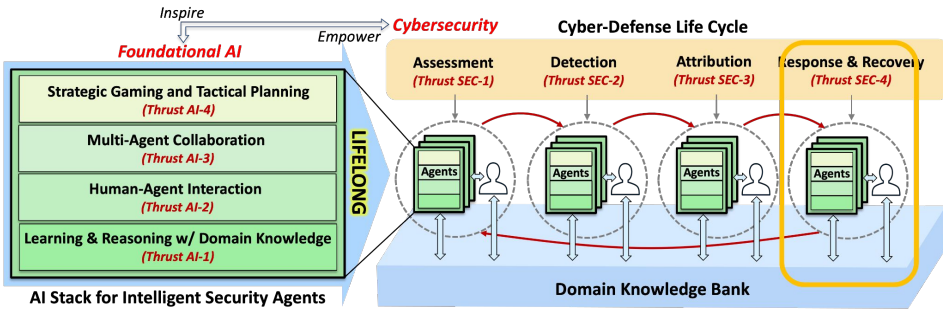
[1] Faculty, [2] UC Santa Barbara, Computer Science

Thrust SEC-4: Intelligent Agents for Response and Recovery

Research Thread 1: Identification of rare events and new attack plan

Project Overview

- Multi-agent anomaly detection-response approach for past, ongoing, and future attacks.
- Uses knowledge collected on both real-world and simulated worst-case and multi-step attacks, as well as adversary models (Thrust AI-4).
- Detectors are instances of agents described in SEC-2, updated by lifelong learning approach from AI-1.
- Incorporates attack graphs from SEC-1 and SEC-3.
- Phase 1 (Jul 23 - Jun 25) goals:
  - Develop multi-detector framework to investigate anomalies caused by each attack step and analyze a log system that records system activity at various levels of detail.



Overview of the ACTION Institute's integrated AI and cybersecurity research.

Challenge

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

Solution

- Following self-supervision, incorporation of human-in-the-loop (AI-2) and continuously learning knowledge base (AI-1, SEC-1) allows for real-time feedback that is scalable.

Achievements this period

- Huang, M., Lee, H., Kundu, A., Chen, X., Mudgerikar, A., Li, N., & Bertino, E. (2023). ARIoTEDef: Adversarially robust iot early defense system based on self-evolution against multi-step attacks. ACM Transactions on Internet of Things.
- Framework/tool to create labeled datasets for multi-step attacks for IoT systems planned for public release in Year 2: IoTDSCreator.

Wenke Lee [1,2], Elisa Bertino [1,3], Ninghui Li [1,3]

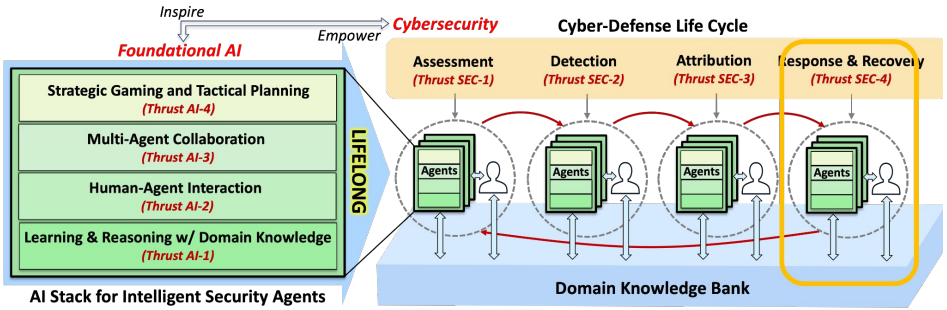
[1] Faculty, [2] Georgia Institute of Technology, Computer Science, [3] Purdue University, Computer Science

Thrust SEC-4: Intelligent Agents for Response and Recovery

Research Thread 2: Automated and robust attack recovery

Project Overview

- How can LLMs be used to model and automate recovery processes in computer networks?
- New forensic analysis tools to significantly boost the cybersecurity of machine learning systems by enabling post-attack traceback that scalably and reliably identifies the source of attacks.
- Phase 1 (Jul 23 - Jun 25) goals:
  - Leverage examples of the existing “security actuators” to train intelligent agents.
  - Agents use the models to suggest, and explain, an action plan to disrupt an attack and improve the security posture of a target.



Overview of the ACTION Institute’s integrated AI and cybersecurity research.

Challenge

- Current approaches are not robust against sophisticated adversarial manipulations.

Solution

- AI stack, integrated with SEC-1 and SEC-2 will improve the state of agents’ robustness.

Future plans

- On-time completion of this project will encompass efforts during next reporting period.
- During overlapping phase 2 (Jul 24 - Jun 26), ACTION researchers will extend LLM-based automation to learn models of security flaws and automatically generate patches.

Wenke Lee [1,2], Nick Feamster [1,3], Bo Li [1,3], Elisa Bertino [1,4], Giovanni Vigna [1,5], David Wagner [1,6], Gang Wang [1,7], Radha Poovendran [1,8], Dongyan Xu [1,4]

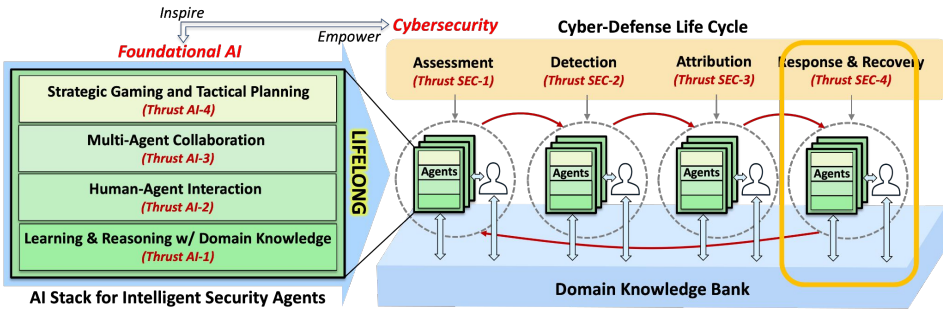
[1] Faculty, [2] Georgia Institute of Technology, Computer Science, [3] University of Chicago, Computer Science, [4] Purdue University, Computer Science, [5] UC Santa Barbara, Computer Science, [6] UC Berkeley, Computer Science, [7] University of Illinois, Urbana-Champaign, Computer Science, [8] University of Washington, Electrical and Computer Engineering

Thrust SEC-4: Intelligent Agents for Response and Recovery

Research Thread 3: Efficient and adaptive response strategies

Project Overview

- Devise solutions for efficient response and recovery to contain the impact of attacks by AI-enabled adversaries.
- Build on our research in foundational AI in AI-2, AI-3, and AI-4.
- SEC-2 and SEC-3 provide foundation for testing and preparation of unforeseen responses.



Overview of the ACTION Institute's integrated AI and cybersecurity research.

Challenge

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

Solution

- Incremental development and integration of all prior thrusts will provide environment for novel solution.

Future plans

- Phase 1 (Jul 23 - Jun 25) goal: 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.
- Phase 2 (Jul 24 - Jun 26) goals: extend our solutions to consider sophisticated attacks that impact multiple networks/sectors, considering interdependencies between the networks and even non-technical factors.

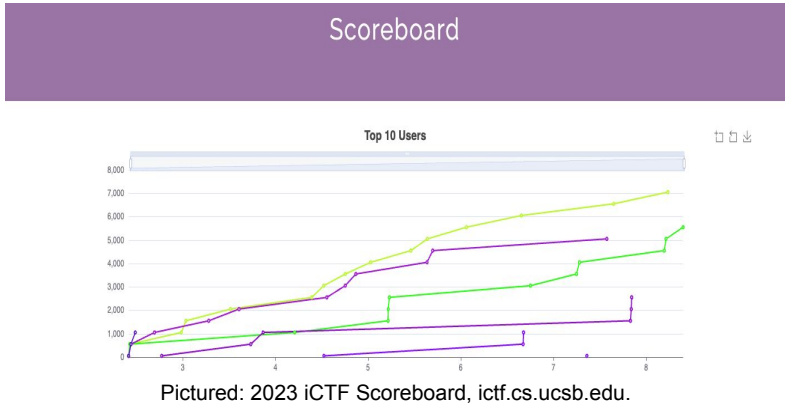
Wenke Lee [1,2], Nick Feamster [1,3], Bo Li [1,3], Elisa Bertino [1,4], Giovanni Vigna [1,5], David Wagner [1,6], Gang Wang [1,7], Radha Poovendran [1,8], Dongyan Xu [1,4]

[1] Faculty, [2] Georgia Institute of Technology, Computer Science, [3] University of Chicago, Computer Science, [4] Purdue University, Computer Science, [5] UC Santa Barbara, Computer Science, [6] UC Berkeley, Computer Science, [7] University of Illinois, Urbana-Champaign, Computer Science, [8] University of Washington, Electrical and Computer Engineering

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 1 goals: launch iCTF, sponsored by ACTION.
- iCTF can be re-packaged for education and workforce development (potential for industry collaboration).



Pictured: 2023 iCTF Scoreboard, ictf.cs.ucsb.edu.

Challenge

- High school students are unprepared for the demands of CTF, or they are hesitant to participate.

Solution

- Increased outreach in the schools.
- Online videos/tutorials to help students through challenges.

Achievements this period

- Collaboration with Women in Computer Science (WiCS) to promote the event beyond ACTION campuses.
- Data collected from all participation can be used 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.
- 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: WiCS/ACTION Outreach at SB High School, April 2024.

Challenges

- Engaging the students, who are already busy, or might be intimidated by CS concepts.
- Finding the tutors to work in schools.

Solution

- We will seek additional funding to help support the dedicated graduate students, to afford them additional time with the students.

Achievements this period

- 2023: Saturday classes through collaboration with UCSB community outreach programs.
- 2024: Collaboration with Women in Computer Science (WiCS).
- 2024: Connection with Santa Barbara High School.
  - Initial outreach during homeroom (see photo).
  - Concrete, measurable plans for additional outreach among students in 2024-2025.
- Awarded UCSB Chancellor's 2024 Faculty Outreach Grant.

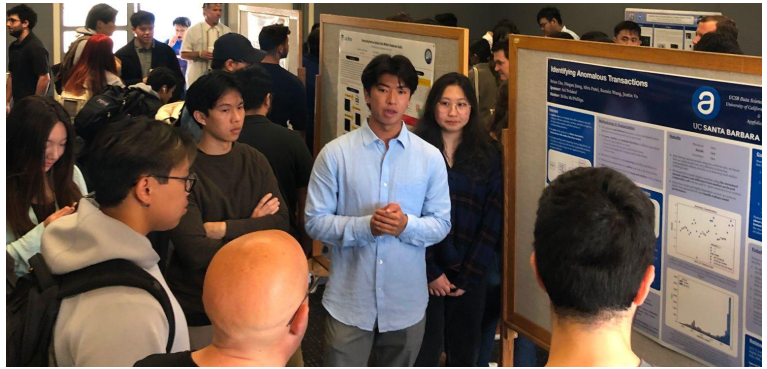
Giovanni Vigna [1,5], Stijn Pletinckx [2,5], Ilya Grishchenko [4,5], Laboni Sarker [2,5], Tim Robinson [3,5]

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

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.
- Provides opportunities for directed research, professional development, and networking for 8 UG and 4 Grads.
- Promotion at ACTION campuses, but primarily at MSIs.
- Focus is on broadening participation and increasing diversity, equity, and inclusion. Goals are at least 50% participation from underrepresented groups.



Pictured: Undergraduate Research Symposium at UC Santa Barbara.

Challenges

- Reaching diversity/participation goals.
- 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 at UCSB.

Achievements this period

- 8 undergraduates recruited. 4 graduate students.
- Students work in teams of 3 or 4 alongside senior-level ACTION graduate student or postdoc. 4 projects planned for summer 2024.
- Secure application portal deployed. More than 100 applications received.
- Help from ACTION Student Advisory Council to select students.

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.