The NSF AI Institute for Agent-based Cyber Threat Intelligence and Operation (ACTION), a National AI Research Institute

Strategic and Implementation Plan, Version 4 (S&IPv4) Program Years 4-5

1 Intellectual Theme **4**
1.1 Goals 5
2 Research Plan **7**
2.1 Overarching Example 10
2.2 Intelligent Security Agents for Cybersecurity 12
2.3 Globally Accessible Testing Environment (GATE) 13
3 Foundational AI Research **15**
3.1 Thrust AI-1: Learning and Reasoning with Domain Knowledge 15
3.1.1 Proposed Research for Years 4-5 16
3.1.2 Planned Research Activities 19
3.2 Thrust AI-2: Human-Agent Interaction 22
3.2.1 Proposed Research for Years 4-5 22
3.2.2 Planned Research Activities 26
3.3 Thrust AI-3: Multi-Agent Collaboration 27
3.3.1 Proposed Research for Years 4-5 28
3.3.2 Planned Research Activities 30
3.4 Thrust AI-4: Strategic Gaming and Tactical Planning 31
3.4.1 Proposed Research for Years 4-5 32
3.4.2 Planned Research Activities 33

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.

3.5 Timeline of Activities 35
4 Use-Inspired Research 36
4.1 Thrust SEC-1: Intelligent Agents for Threat/Vulnerability Assessment 37
4.1.1 Proposed Research for Years 4-5 37
4.1.2 Planned Research Activities 39
4.1.3 Metrics and Outcomes 41
4.2 Thrust SEC-2: Intelligent Agents for Detection and Understanding 41
4.2.1 Proposed Research for Years 4-5 42
4.2.2 Planned Research Activities 43
4.2.3 Metrics and Integration 44
4.2.4 Reference Applications and GATE Testing 45
4.3 Thrust SEC-3: AI-Driven Cyber Threat Intelligence (CTI) 45
4.3.1 Proposed Research for Years 4-5 45
4.3.2 Planned Research Activities 46
4.4 Thrust SEC-4: Intelligent Agents for Response and Recovery 48
4.4.1 Proposed Research for Years 4-5 48
4.4.2 Planned Research Activities 50
4.4.3 Metrics and Integration 52
4.5 Timeline of Activities 53
5 Research Risks and Mitigation 53
6 Education and Workforce Development 57
6.1 K-12 & Community 58
6.2 Undergraduates 61
6.3 Graduate Students 64
6.4 Postdocs and Advanced/Senior PhD Students 66
6.5 Workforce Development 69
6.6 Additional Education-related Projects, Year 4 71
6.7 Assessment and Evaluation 74
6.8 Education Risks and Mitigation 74
7 Integration 75
7.1 ACTION's Integrated Research Approach 76
7.2 Risks to Integration and Mitigation 79
8 Broadening Participation 80
8.1 Broadening Participation Risks & Mitigation 82
9 Collaboration and Knowledge Transfer 85
10 Participation and Access Plan 88
10.1 Vision 88

10.2 Definitions 88
10.3 Hiring Practices 89
10.4 Decision-making and Governance 89
10.5 Societal Impact 90
10.6 Monitoring, Evaluation, and Assessment 90
11 Management 91
11.1 Organization 92
11.2 Research Prioritization and People Management 96
11.3 Key Personnel 98
12 Intellectual Property 100
12.1 Publications, Datasets, and Code 100
12.2 Commercialization and Licensing 101
12.3 Research Data 101
12.4 Evaluation Data 102
13 Ethics 103
14 Continuity 106
15 References Cited 107

1 Intellectual Theme

The cybersecurity threat landscape is constantly changing. Cyber-attacks continue to evolve, introducing new adversaries, targets, vulnerabilities, strategies, and tactics. As a result, cybersecurity is a major concern to any system, service, or infrastructure that has a cyber component: from schools to hospitals, from power plants to military bases.

A typical cyber-defense life cycle includes risk assessment and prevention, detection, attribution, and response and recovery. Today, these tasks remain heavily dependent on human defenders' skills, intuitions, and experience, and their effectiveness and efficiency do not meet the requirements of next-generation cybersecurity in terms of speed, scale, and ability to operate in an uncertain world. The threat surface is expanding beyond traditional digital infrastructure: intelligent agents are increasingly deployed in or connected to physical systems, from industrial control environments to autonomous robotic platforms, creating new attack surfaces and new requirements for AI-driven defense.

The Institute for Agent-based Cyber Threat Intelligence and Operation (ACTION) continues its 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.

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. A primary deliverable of Year 4 is the formal definition and documentation of this AI stack, consolidating the foundational contributions of the institute's four AI thrusts into a coherent, reusable architecture that industry partners and the broader research community can build upon.

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.

1.1 Goals

In its first five years of operation, the ACTION Institute aims to achieve three main goals, including impacts on research, education, and technology transfer.

1. Support the development of security solutions based on intelligent security agents that can operate at the scale and speed necessary to address current and future threats to our critical infrastructure. This will be achieved via integrated, collaborative research across multiple, leading universities that synergizes novel foundational AI concepts and use-inspired approaches to cybersecurity. Results from interdisciplinary research spanning Computer Science, Statistics, and Engineering will expand the foundations of AI and enable high automation, scalability, and cost-effectiveness in both proactive and reactive cyber-defense (see Figure 1). 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. In Year 4, this virtuous circle extends explicitly to physical AI and robotic systems, as ACTION's research agenda broadens to include the security of AI-controlled physical infrastructure.

2. Educate students and the workforce about how AI and applications based on intelligent security agents can be used to protect the nation's critical infrastructure against attacks. To this end, the ACTION Institute's research efforts will be complemented by innovation in education (K-12, undergraduate, graduate, and post-graduate) and workforce development tools, which will be made available to the community. The ACTION Institute will expand existing offerings such as courses on AI and security by creating integrated coursework that focuses on how AI can help solve security problems, and, at the same time, how a deep understanding of security issues can push the limits of what can be done with AI today, inspiring new directions in research. The Institute will also develop novel approaches to education in AI and security. In particular, it will develop cybersecurity competitions centered around AI that involve high school and undergraduate students, creating opportunities to learn integrated AI and cybersecurity concepts in a highly motivating setting. The Agentish platform, developed in Year 3, introduces agentic programming as a core educational

competency alongside traditional CTF skills. To reach a population that goes beyond the constituents of the Institute's member institutions, we plan to use the "network of networks" approach, in which groups with shared interests and focus operating at different locations collaborate and strategize to reach a broad audience. By making the educational tools widely available using MOOCs, open-source software, and open datasets, the Institute will make educational opportunities about AI and cybersecurity available to the security workforce. The institute will reach out to the workforce to publicize these efforts using local workshops and other educational events, which will be open to the public and not limited to the student population.

3. Transfer concepts, approaches, and technologies to industry stakeholders so that they can become commercial products deployed to protect organizations, enterprises, and governments. The ACTION Institute will design an integrated AI stack that supports the development of intelligent security agents, and will introduce a cyber range, the Globally Accessible Testing Environment (GATE), for the testing and demonstration of said agents. By showing how advanced AI capabilities can be composed to perform security tasks, the ACTION Institute will provide blueprints for the technology industry stakeholders so that they can develop commercial-grade solutions that incorporate the findings of the ACTION Institute researchers. This transfer will be fostered by regular interaction with industry partners, with the goal of extending the impact of the Institute beyond research and education. The emergence of Virtue AI and Artiphishell as startups, and the public release of ACTION tools such as TOUCAN and CVE-GENIE, show this transfer pipeline at work. At the same time, the creation of a series of demo agent-based applications that leverage the AI stack and run on the GATE infrastructure will provide focus and integration among the various research and educational activities of the Institute. By fostering collaboration among the member organizations and with external industry partners, the ACTION Institute will become a nexus for the AI and cybersecurity communities at both the academic and industrial/commercial levels.

Our team is well-poised to impact education and broadening participation in computing. Leading our educational efforts, Dr. Radha Poovendran is well-known for his innovative contributions to classroom learning, including the development of an AI-powered instructional chatbot and a novel graduate course on autonomous vulnerability analysis. Dr. Giovanni Vigna has introduced novel methods for teaching computer security through the deployment of Capture the Flag exercises at the

university and high school level, and has led the development of the Agentish platform for agentic programming education. Dr. Dawn Song has considerable experience in the development of Massive Open Online Courses, with previous MOOCs on LLM Agents reaching more than 23,000 learners globally; a new AI+Security MOOC is planned for Year 4. Moreover, Institute faculty have established ties with public education in communities near ACTION campuses in Chicago, Atlanta, and Santa Barbara, and have expanded outreach to school systems across the Hampton Roads region of Virginia through our partnership with Norfolk State University. Dr. Wenke Lee, who has substantial experience in leading educational activities, will direct outreach and work closely with education coordinators at partner institutions.

By creating a network of collaboration among AI and cybersecurity researchers, the ACTION Institute will connect an innovative community that is greater than the sum of its parts, capable of research that will change the way we design and implement cybersecurity solutions.

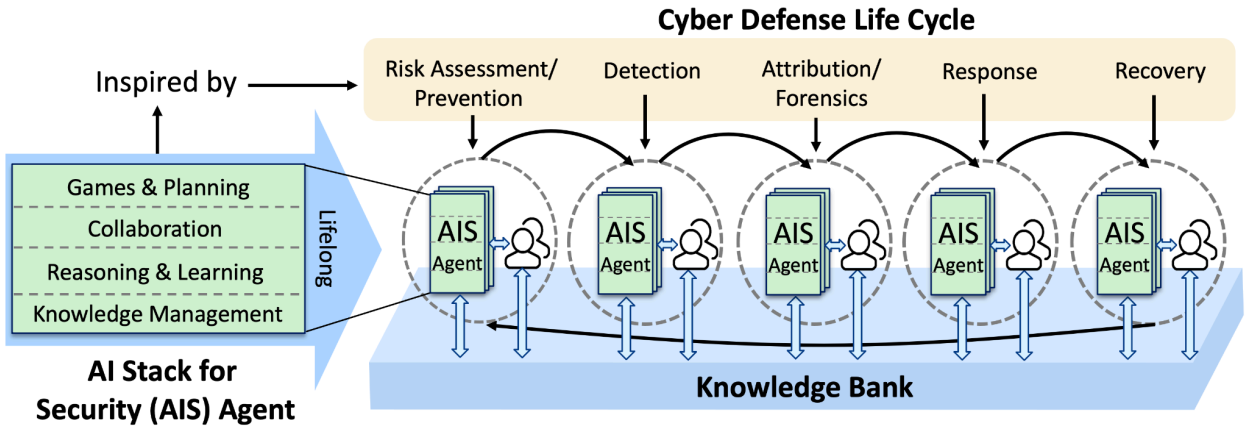


Figure 1: Overview of the ACTION Institute's integrated AI and cybersecurity research.

2 Research Plan

Since its launch in June 2023, ACTION has established itself as a capable and productive research institute, with over 300 grant-acknowledged manuscripts, an h-index of 44, and a growing portfolio of open-source tools deployed by industry and government partners. As the institute enters Years 4 and 5, we intend to maintain our momentum and continue with the research plan visualized in Figure 1. The ACTION Institute's

research plan is defined by the axes of foundational AI and cybersecurity. Each axis contains four critical elements, and each of these elements maps to one of eight research thrusts. Along the axis of artificial intelligence, we focus on four core questions concerning the capabilities of intelligent security agents:

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

Answering these questions requires overcoming certain challenges, including the limitations of current AI approaches and techniques. The subsequent investigation will lead to the creation of an AI stack that empowers intelligent agents in novel applications for cybersecurity. The questions therefore frame four of ACTION's research thrusts:

- Thrust AI-1: Learning and Reasoning with Domain Knowledge
- Thrust AI-2: Human-Agent Interaction
- Thrust AI-3: Multi-Agent Collaboration
- Thrust AI-4: Strategic Gaming and Tactical Planning

The cybersecurity axis of our research plan maps a use-inspired thrust to each of the four phases of the cyber-defense life cycle: assess, detect, attribute, and respond/recover. Each thrust examines the impact of intelligent agents on a particular phase.

- 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 and sharing of information.

- 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.

The space defined by the axes of foundational AI and cybersecurity will be the testing ground for our Globally Accessible Testing Environment (GATE), a novel infrastructure for deploying and experimenting with intelligent agents. GATE, which is further described below in Section 2.3, allows for testing integrated prototypes with increasing levels of sophistication. Prototypes bring ACTION researchers closer to our industrial partners, where collaboration and knowledge transfer then lead to testing of proposed research ideas in real-world computer systems. As the possibilities for AI in real-world security operations broaden, there will come new questions concerning the limits of the technology. These questions will further refine our research.

So it is that our research activities will be continually intertwined during the next two years of ACTION and beyond. The creation of an AI stack will impact aspects of the cybersecurity lifecycle. Advances by AI agents in cybersecurity will recursively inspire improvements in the AI stack. Those improvements in AI will then lead to further improvements in cybersecurity, and a virtuous circle will continue.

In the following, we first introduce an overarching example in Section 2.1, which will be used as a reference to showcase the novel approaches to cybersecurity brought forward by the proposed research. In Section 2.2 we detail the proposed impact of our plan for intelligent agents. We then break down each of our AI and use-inspired thrusts into smaller threads and highlight proposed projects that both connect our research threads and serve as milestones for program evaluation. Section 3 discusses our foundational AI thrusts, which, when combined, realize an AI stack providing novel capabilities for knowledge learning and modeling, human-agent interaction, multi-agent collaboration, and strategic gaming and tactical planning for intelligent security agents. Section 4 describes the use-inspired research thrusts that leverage the AI stack to create a framework that addresses the assessment of the security posture of a system, the detection of sophisticated attacks, the ability to attribute malicious actions to adversaries, and the execution of response and recovery steps.

There are risks that might hinder the progress of the team, and create roadblocks in both research and education activities. These risks are described in more detail in Section 5, along with planned mitigations.

Given the proliferation of AI, there is an immediate need for a focused, structured plan for intelligent AI agents in the cybersecurity landscape. ACTION's multi-year research collaboration of a large group of top researchers across multiple institutions and disciplines tackling integrated research thrusts provides this plan. ACTION's core AI researchers have been at the forefront of expanding the foundations of AI, machine learning, game theory, and computer security.

The ACTION team also includes leading researchers in cybersecurity.

2.1 Overarching Example

We will use an overarching example to exemplify how our advances in foundational AI and our use-inspired research will change the way in which intelligent agents collaborate and interact with humans to improve cybersecurity. The attacks described hereinafter are inspired by the PIs' real-world experience with sophisticated cyber-attacks, such as the recent SolarWinds supply-chain attack [1] and the large-scale exploitation of the Log4j vulnerability [2]. We consider a scenario in which a nation-state actor is performing a multi-pronged attack against the critical infrastructure of a (fictional) US city (see Figure 2). More precisely, the operation's goal is to create uncertainty and chaos by shutting down the power and water distribution

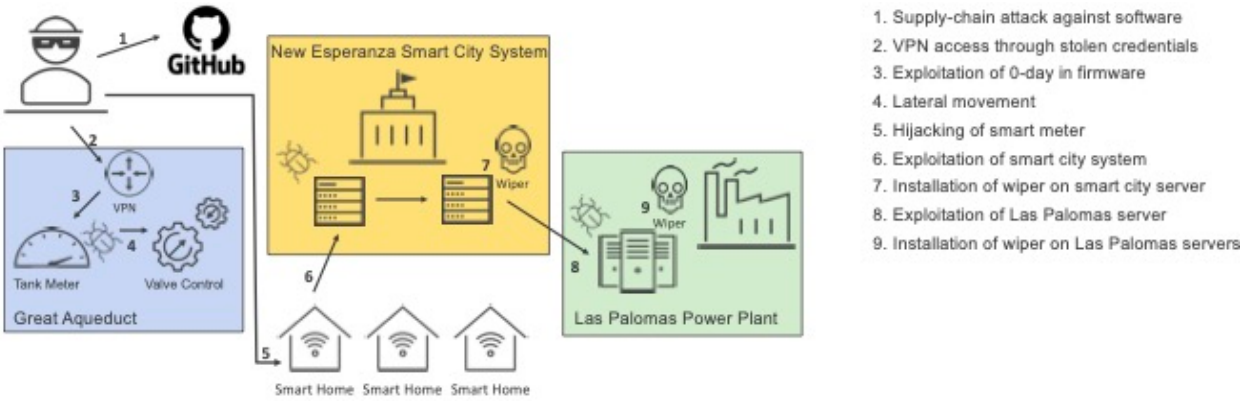


Figure 2: Steps of the overarching attack example. The attackers first perform a supply-chain attack (1), followed by the compromise of the Great Aqueduct (2, 3, 4). The attack against the Smart City system (5, 6, 7) is then used as a bridgehead to attack the Las Palomas power plant (8, 9).

infrastructure of New Esperanza. To this end, the attackers focus on a specific water distribution system, the Great Aqueduct, and a central power production and distribution system, the Las Palomas power plant. The systems that control the operations of the Great Aqueduct and the Las Palomas plant are integrated with New

Esperanza's smart city system, whose open-source software monitors and distributes power, water, and other services to the city's neighborhoods.

The nation-state actor begins its operation by gathering intelligence about the targets of the attacks. Early in the process, they identify the open-source software used by the smart city system as a possible target for a supply-chain attack and are able to contribute a vulnerable software component to the project using false identities. This attack goes undetected because the introduced vulnerability is complex and logic-based, and current vulnerability analysis systems are unable to detect it.

The attackers use a set of credentials obtained from underground forums to successfully connect to the VPN of the Great Aqueduct system. The initial access is limited to a set of water tank level monitors. However, the attackers are able to quickly identify a previously unknown flaw in the firmware of the tank-level monitors and use the compromised monitors to pivot to a valve control system that directly regulates the distribution of water. The privilege escalation is not detected as there are no security components inside the water tank level monitors. Even though the communication between the level monitors and the valve control system is suspicious, there is no deployed anomaly detection system that is able to identify the unusual connection, and, as a result, the lateral movement is not identified.

The second attack phase focuses on New Esperanza's smart city system. By hijacking the connection between the smart city system and one of the smart meters deployed in a neighborhood, the attackers exploit the vulnerability previously introduced in the open-source software used by the smart city system to obtain administrative access to the main server and upload a wiper malware component. The exploitation of the 0-day vulnerability is not detected because it is unknown at the time of the attack.

Access to the smart city system allows the nation-state attackers to scan the network of the Las Palomas power plant. The attackers identify an unpatched host that is immediately compromised and then used as a bridgehead to distribute the wiper malware to multiple hosts in the plant network. While the scanning attack is detected, it is not prioritized and brought to the attention of a human operator. The unpatched host exploitation happens through an encrypted connection and the power plant network intrusion detection system is not able to observe the attack.

Once the attackers realize that the various components of the attack are in place, they activate the wipers, erasing the file system and boot record of each server. As a result, the power plant ceases operations, and the smart city system cannot be controlled. At

the same time, the malware that controls the water plant valve is activated, shutting down the aqueduct and blocking the flow of water to the city. Since the smart city system is incapacitated, it cannot reroute the water and power resources necessary to deal with the failure of the Great Aqueduct and the shutdown of the Las Palomas power plant. As a result, the city is paralyzed and chaos ensues.

2.2 Intelligent Security Agents for Cybersecurity

The ACTION Institute brings forward innovation in AI and its application to cybersecurity that will protect critical infrastructure from sophisticated attacks such as the one experienced by New Esperanza. Through three years of research, ACTION has developed the foundational agent capabilities that, when integrated, are capable of addressing each phase of the attack described above. The institute's Year 4 agenda focuses on deepening these capabilities and integrating them into a coherent, operational AI stack that can be validated end-to-end in GATE.

Our Thrust AI-1 continues to develop novel ways to learn and reason about domain knowledge, which are leveraged by Thrust SEC-1 to implement Knowledge Agents. These agents are able to discover important facts about the infrastructure, such as the fact that the smart city network could be used to access the power plant's server hosts. This knowledge is then used by the Unit Assessment Agents and System Assessment Agents of Thrust SEC-1, whose task is to identify flaws in components and systems. These agents are designed to identify the kinds of logic-based supply-chain vulnerabilities and firmware flaws that allowed the New Esperanza attack to proceed undetected.

The Detection Agents outlined in Thrust SEC-2 collaborate using the multi-agent paradigms provided by Thrust AI-3 to gather information and capture models of the normal operation of the system, which are then stored in the domain knowledge bank designed by Thrust AI-1. These models are used as a basis to detect anomalous events, such as the communication between the water tank monitor and the water valve control system, which are brought to the attention of the human user using the human-agent interaction approaches outlined in Thrust AI-2. As agents are increasingly deployed in physical environments, including industrial control systems and robotic platforms, the Detection Agents must also account for the cyber-physical attack surface, an expansion that GATE's growing cyber-physical infrastructure is designed to support.

The attackers' actions observed during the information-gathering phase will be compared to the event traces generated by the simulation of malicious multi-agent scenarios enabled by the research of Thrust SEC-3, with the goal of determining the identity of the attackers and their motives. The modeling of the possible adversaries relies on the multi-agent models provided by Thrust AI-3 and the strategy-game-exploration techniques outlined in Thrust AI-4. For example, the initial steps of the attackers would allow the Attribution Agents to identify the attacker as a nation-state whose goal is to disrupt essential services. As a result, the operators of the network would be alerted of the destructive intent of the attack and prepare redundant mechanisms for power and water distribution.

The information collected by the detection and attribution agents of Thrusts SEC-2 and SEC-3 is then used to drive the action of Response and Recovery (R&R) Agents, which are the focus of Thrust SEC-4. These R&R Agents use the action-planning primitives provided by Thrust AI-4 as well as the agent-agent collaboration mechanisms provided by Thrust AI-3 to define the best course of action in containing the threat and reducing its impact, which will include humans in the decision-making loop using the primitives of Thrust AI-2. Using the learning and reasoning primitives provided by Thrust AI-1 as well as the multi-agent cooperation of Thrust AI-3, the R&R Agents will cooperate with the Unit and System Assessment Agents to make the network more resilient to future attacks. Critically, the R&R Agents must also recover from compromised memory and reasoning states, ensuring that agents whose knowledge was poisoned during an attack can be restored to trustworthy operation — a capability under active development in Year 4.

2.3 Globally Accessible Testing Environment (GATE)

As the central venue for research integration and evaluation, ACTION has developed GATE as a virtual integrated testbed that brings together the tools, agents, datasets, and scenarios produced across all eight research thrusts. Entering Year 4, GATE has a functioning architecture with defined interfaces, a growing library of attack and defense agents, and demonstrated cyber-physical evaluation capabilities including embodied robotic AI scenarios.

Faculty associated with each research thrust develop specific test environments, datasets, software systems, and libraries to evaluate the proposed models and approaches. GATE integrates these environments to run large-scale and comprehensive experiments that evaluate all facets of the Institute's research in realistic settings. Each thrust's environment defines a data model and software interface, and standardized

protocols and APIs allow these environments to be linked to form the integrated testbed. All authorized researchers and practitioners have access to GATE and are able to evaluate scenarios that involve capabilities provided by multiple thrusts.

GATE's design is intentionally flexible, integrating a spectrum of simulation and emulation capabilities. It incorporates digital twins of real-world critical infrastructure, including industrial control systems and autonomous robotic systems, for high-fidelity security evaluations. GATE is engineered to bridge with other specialized testbeds, such as SPHERE [3], enabling complex hardware-in-the-loop and large-scale emulation experiments that are critical for validating research against real-world conditions.

In Year 3, GATE matured from a planned infrastructure into an active research platform. Key results include new frameworks for least-privilege LLM agents and for finding adversarial failures in embodied robotic AI. Full details are provided in the Year 3 annual report.

In Year 4, GATE will pursue two primary directions. The first deepens its investment in physical AI and robotic system security, with a specific focus on LLM-controlled robots. The team plans to deploy the attack, defense, and benchmark agents developed in Year 3 directly into GATE's infrastructure, expanding its cyber-physical evaluation capabilities and making these tools available to practitioners and educators who lack access to physical robots for security testing. The second direction extends GATE into reverse engineering and binary analysis, integrating two complementary components: a Codex-based reverse engineering agent that serves as a reference autonomous analyst capable of solving CTF-style problems, and an enhanced skill-based reverse engineering tool that structures reverse engineering into modular skills for controlled experimentation and benchmarking.

Two key features of GATE are user-friendly experiment set-up and always up-to-date configuration. All libraries and datasets from all thrusts are accessed remotely through web-based APIs. Researchers use a web-based user interface to specify and run experiments via GATE, remotely calling relevant APIs and datasets according to the experiment's requirements. GATE maintains a repository for current and historical libraries and APIs, datasets, and documents, as well as build and configure scripts to facilitate repeatable experiments.

In Year 4, ACTION will establish a central GATE Collaboration Portal to serve as a hub for sharing experiment artifacts, datasets, and agents, ensuring reproducibility and accelerating knowledge sharing across the institute. GATE will also serve our

educational mission through a distributable lightweight version, a "GATE-in-a-box," that packages research tools and scenarios into accessible curricula for use by institutions that lack the resources to run a full cyber range. An integrated CTF platform will directly translate our research into hands-on learning experiences, improving workforce development and providing a practical means to disseminate our technologies to a broader community of students, researchers, and practitioners.

At each annual project meeting, each thrust will report status on the availability of its libraries and datasets and their integration into GATE. Each project meeting will include a demonstration session on thrust-specific capabilities, as well as joint experiments using GATE.

3 Foundational AI Research

The ACTION Institute pursues foundational AI research to create an integrated AI stack (Figure 1) for intelligent security agents, with (AI-1) learning and reasoning with domain knowledge, (AI-2) human-agent interaction, (AI-3) multi-agent collaboration, and (AI-4) strategic gaming and tactical planning layers, as well as domain knowledge bases to be accessed and managed by the agents. Our research has been, and will continue to be, conducted according to four thrusts, associated with the four layers of our AI stack. Below, for each AI thrust, we highlight the main challenges presented by the status quo, the proposed contributions of our research over the next two years of ACTION's operation, and our pending research projects that will help us realize this vision.

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

Learning and reasoning are two essential AI primitives for intelligent security agents, and these primitives can be empowered with domain knowledge to improve their effectiveness and resilience. To realize knowledge-based learning and reasoning, we propose the following research efforts: develop scalable reasoning methods on dynamic graphs, integrate reasoning across knowledge graphs and LLMs, provide continual anomaly detection and learning, and advance veridical learning and reasoning with end-to-end quality control.

Challenges:

- Current explanation methods (white box or black box) are static and not powerful enough for explaining and reasoning about large dynamic graphs.
- Current AI techniques lack the ability to integrate domain knowledge and the inference ability of logical reasoning for prediction and decision-making tasks.
- Current approaches lack the scalability and lifelong learning ability for anomaly detection.
- Current AI agent pipelines lack principled end-to-end quality control frameworks to ensure that learning and reasoning outputs are reproducible, stable, and trustworthy across the data science life-cycle.

To address these challenges, our proposed research efforts will lay the core theoretical foundation of explaining and reasoning and will remain one of the differentiating themes of ACTION.

3.1.1 Proposed Research for Years 4-5

Thread 1: Reasoning on graphs. Graphs are fundamental structures for representing complex relationships in data, capturing entities as nodes and their interactions as edges. Graphs model interconnected information in ways that traditional data formats cannot. Reasoning over graphs is essential for drawing inferences, discovering hidden patterns, predicting future connections, and answering complex queries that involve chains of relationships. As datasets grow increasingly relational and high-dimensional, the ability to reason with graphs becomes critical for advancing machine learning in domains where context and structure define meaning.

However, reasoning with graphs poses several significant challenges. Graphs often involve heterogeneous, incomplete, and noisy data, making it difficult to extract reliable insights. Scaling reasoning methods to large/dynamic graphs requires algorithms that are both efficient and adaptive. Capturing long-range and hidden dependencies adds further complexity. These challenges demand continued innovation in areas like graph neural networks, symbolic reasoning, and probabilistic inference to create systems capable of robust, scalable, and transparent reasoning over complex relational data.

In cybersecurity, reasoning over graphs is increasingly vital as threats often emerge from complex, multi-step interactions across vast networks of devices, users, and systems. Attack detection, vulnerability analysis, and fraud prevention all benefit from modeling these relationships as graphs, whether through tracing intrusion paths, mapping dependencies, or identifying anomalous patterns in communication flows.

AI-powered graph reasoning enables systems to move beyond simple pattern matching, allowing them to infer hidden attack vectors, predict potential breaches, and adapt defenses in real time. Yet, this integration of graph reasoning and AI in cybersecurity faces challenges, including handling incomplete or obfuscated data from adversaries, ensuring scalability in high-throughput environments, and maintaining robustness against adversarial manipulation. The intersection of AI, graph reasoning, and cybersecurity thus represents a critical frontier, where understanding relational structure is key to building intelligent, proactive defenses against increasingly sophisticated threats.

Thread 2: Integrated reasoning across knowledge graphs and LLMs. Integrating domain knowledge into large language models (LLMs) through knowledge graphs (KGs) offers a promising path toward more grounded, accurate, and explainable AI systems. While LLMs excel at capturing statistical patterns from massive text corpora, they often lack the explicit, structured understanding that domain-specific knowledge graphs provide. Knowledge graphs encode curated relationships, ontologies, and facts about entities in a way that reflects expert knowledge, which can complement LLMs by anchoring their generative abilities to verified information. This integration allows for more reliable reasoning in complex tasks where factual accuracy and logical consistency are critical. By combining LLMs' language capabilities with the relational structure of KGs, systems can both "know" and "reason," dynamically pulling in relevant knowledge and applying it to unseen problems.

However, achieving seamless reasoning across LLMs and KGs poses significant challenges. Aligning the unstructured representations learned by LLMs with the symbolic, discrete structures of knowledge graphs requires sophisticated methods for knowledge retrieval, entity linking, and relation grounding. Furthermore, ensuring that the integrated system respects domain constraints such as logical rules, hierarchies, or causal relationships requires new approaches to controlled reasoning and consistency enforcement. Addressing these challenges is key to building hybrid AI systems that can reason transparently and robustly, combining the breadth and flexibility of language models with the depth and precision of human-curated knowledge.

The integration of knowledge graphs (KGs) and large language models (LLMs) is becoming increasingly important for advancing threat analysis, vulnerability discovery, and automated defense mechanisms. Cybersecurity operates in complex, rapidly evolving knowledge domains: attack patterns, system dependencies, threat actor behaviors, and protocol vulnerabilities, which are naturally suited to structured representation in knowledge graphs. These KGs encode rich, domain-specific

relationships that reflect years of expert research, enabling systems to capture causal chains (e.g., from exploit to breach) and organizational context (e.g., asset criticality). When combined with LLMs, which can process unstructured reports, advisories, and real-time logs, these hybrid systems open the door to reasoning across both curated knowledge and live data, enabling more accurate threat detection, explanation, and response.

Thread 3: Lifelong or continual anomaly detection and learning. Detecting intrusions and attacks is an important task in security. Anomaly detection is the common technology used for the task. A parallel task is to classify the detected anomalies (intrusions or attacks) into known types so that they can be dealt with using their corresponding measures. For this task, we normally use a classification model. However, the current model building approaches for both anomaly detection and attack classification are offline methods. They learn a fixed model from a given training dataset.

However, in any security application, new types of intrusion and attacks appear constantly, and existing models may also make mistakes because, due to limited training data, they may not be well-trained. It is highly desirable that the detection model is a lifelong or continual learning model that is not frozen and can be updated continually or incrementally after deployment. For example, if a data point is detected as an anomaly, but it is not, as judged by a human expert, then the data point should be used to improve the detection model online so that it will make fewer detection mistakes in the future. Note that we need humans-in-the-loop to ensure the system learns correctly.

The attack classification model should also be a lifelong/continual learning model. For example, if an anomaly is detected and confirmed by the human expert to be a new type of attack, then the attack classification model should incrementally learn the new class of attack so that in the future the model can classify all the past and new classes/types of attacks.

Thread 4: Veridical Learning and Reasoning with End-to-End Quality Control. It is highly desirable for an agent's sensing-reasoning pipeline to have end-to-end quality control. To this end, we will expand Dr. Yu's Predictability, Computability, and Stability (PCS) framework for Veridical Data Science [4]. More specifically, Dr. Yu's prior work on PCS [4] stipulates a well-documented and reproducible data science life-cycle (DSLC), from problem formulation to data collection and choice, data cleaning and curation, to data visualization, algorithm development, and post-hoc analysis with conclusions. For

agents in cybersecurity, our significantly enhanced PCS framework will enable responsible, reliable, and reproducible learning and reasoning on diverse data (for learning) and knowledge (for reasoning) in the cybersecurity domain.

Specifically, we will collect attack and threat data from both the real world and our attack simulation platform (see Section 2.3) and follow the PCS principles to compare the predicted and actual outcomes, thus providing comprehensive quality control and stability analysis for our learning and reasoning pipeline. Furthermore, we will use the PCS framework as another way of knowledge discovery based on its capability of hypothesis generation about security threats, adversarial behaviors, and their relationships, to uncover non-obvious knowledge of cybersecurity functions. This will further enrich the agents' knowledge for the learning and reasoning pipeline.

In Year 4, we will focus on three directions: adapting PCS for intrusion and malware detection use cases; leveraging LLMs to build lifelong learning systems continuously validated through PCS, with reward functions that integrate human feedback; and advancing train-time attack and defense methods, including Shapley value optimization to identify data poisoning and privacy-preserving techniques for robust model sharing.

Metrics: For this thread, we will measure the safety of the built learning-reasoning frameworks at different stages, such as training, knowledge integration, and fine-tuning, to evaluate the security improvement of the framework and validate our training and data and knowledge collection processes.

3.1.2 Planned Research Activities

To meet the challenges of Thrust AI-1, we will undertake the following projects during Year 4:

Project: Self-Motivated and Continuous Learning Agents for Security Applications

Leads: Bing Liu (UIC) and Xiaoming Huo (Georgia Tech)

Abstract: This project aims to develop autonomous AI agents that can perform continual and self-motivated learning in open-world environments, with a particular focus on security applications. Specifically, these agents can identify emerging threats and attacks (tasks), acquire ground-truth information and data through interactions

with human experts and/or other knowledgeable agents, and incrementally learn the new tasks online (on the fly) within their operational contexts.

Outcomes: (1) One or two research papers documenting the proposed algorithms. We will submit them to top machine learning and/or security conferences. (2) The implemented software will be uploaded to GitHub. (3) A theoretical framework for solving the proposed problem.

Primary Thrust: AI-1: Continuous/lifelong detection and learning in the open world environment

Secondary Thrust(s): AI-2: Agent-human interaction and collaboration. SEC-2: Intelligent agents for threat detection.

Project: Assessing and Mitigating Inherited Vulnerabilities in Multi-Modal Foundation Models for Agentic Defense of Critical Infrastructure

Leads: Elisa Bertino (Purdue University) and Sewoong Oh (University of Washington).

Abstract: In the Year 4 scenario (Section 4), a nation-state adversary conducts coordinated attacks against U.S. chemical and water-processing plants by exploiting similarities in their IT infrastructures. Intelligent security agents protect these facilities by continuously assessing security posture, detecting attacks, sharing threat intelligence, and orchestrating response and recovery actions (SEC-1 through SEC-4). A growing number of these agents rely on multi-modal foundation models (e.g., vision-language models (VLMs)), to interpret sensor data, analyze logs, reason over operator instructions, and plan defensive actions. These models are pretrained on large public datasets and then adapted to specific tasks using limited, often proprietary data. While concealing pretrained backbones and downstream models is commonly assumed to improve robustness, it is unclear whether adversarial vulnerabilities in pretrained multi-modal models can propagate to downstream security agents under realistic black-box conditions. This project investigates whether such inherited vulnerabilities undermine the effectiveness of agentic defense across multiple plants with similar infrastructures, thereby enabling scalable, repeatable attacks, the very threat envisioned in the Year 4 scenario.

Integration with ACTION AI Stack and GATE: The proposed research integrates with ACTION's AI stack by strengthening learning (AI-1), collaboration (AI-3), and reasoning-based planning (AI-4) in security agents. Outcomes will inform how agents

reason about uncertainty in multi-modal inputs and coordinate defensive actions across plants.

Primary Thrust: AI-1

Secondary Thrust(s): AI-3 and AI-4. SEC-1, SEC-2, SEC-3, SEC-4.

Project: Secure Agents for Cyber and Patient Interactions

Leads: Bin Yu (UC Berkeley) and Bo Li (University of Illinois Urbana-Champaign)

Project Description: The increasing integration of large language models (LLMs) and AI agents in healthcare presents both tremendous opportunities and serious risks. This project aims to develop robust frameworks for safe reasoning, trustworthy alignment, and query optimization for LLMs and agents used in patient-facing healthcare scenarios. We focus on both the pre-processing (query understanding) and post-processing (response filtering and explanation) phases of interaction, tackling two core challenges: (1) How to structure patient and caregiver queries to guide LLMs and agents towards medically safe and contextually relevant responses. (2) How to align LLMs and their downstream agents to remain compliant with medical ethics, safety protocols, and regulatory guardrails.

Outcomes: Our developed strategies and optimization approaches will be deployed in ACTION's smart city testbed, enabling secure, realistic evaluation using synthetic patient data, EHR integration, and robust testing.

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

Secondary Thrust(s): AI-3: Multi-agent collaboration. SEC-1: Threat assessment.

Project: Latent-Space Detection Systems

Leads: Ambuj K. Singh (UC Santa Barbara) and David Evans (University of Virginia)

Project Description: This project outlines a Latent Space Detection System designed to counter attacks on critical infrastructure. We aim to develop a novel system that can detect previously unknown threats to critical infrastructure systems by using the

internal representation space of an LLM and interpreting changes as geometric trajectories rather than discrete events, shifting cyber defense from reactive rule-matching to geometric analysis.

Outcomes: This work targets SEC-2 and SEC-4 by automating geometric threat detection and response. We rely on AI-1 to update the latent space, AI-4 to compute defensive steering vectors, and AI-2 to translate these into actionable operator directives.

Primary Thrust: AI-1: Learning and Reasoning with Domain Knowledge.

Secondary Thrust(s): SEC-2, SEC-4, AI-2, AI-4.

3.2 Thrust AI-2: Human-Agent Interaction

Intelligent agents are expected to interact seamlessly with humans. We propose to develop principled methods for effective human-agent interaction according to the following three paradigms: Agent-in-the-loop decision-making, learning and reasoning with humans-in-the-loop, and human-agent teaming.

Challenges:
- Current human-AI systems are weak in supporting humans' responses to agents' recommendations and actions.
- Lack of principled ways to solicit high-level human intelligence (i.e., beyond labels) in a scalable and reliable manner to augment autonomous agents.
- Lack of principled ways to conduct task allocation and information exchange within a team composed of both humans and agents to optimize the effectiveness of the team.

Our three proposed threads map to three paradigms that cover the full spectrum of human-agent interaction. They form a comprehensive and stable research agenda for this thrust, despite rapid advances in areas such as large language models (LLMs).

3.2.1 Proposed Research for Years 4-5

Thread 1: Agent-in-the-Loop Decision-Making. This paradigm reflects the scenarios where humans are the final decision makers in a decision-making task (e.g., detect attacks), while agents provide assistance in the form of direct decision

recommendations or indirect intelligence collection/analysis. We will enhance the effectiveness of agent-in-the-loop decision-making through three perspectives:

1. **Promoting humans' calibrated trust in agents.** In agent-in-the-loop decision-making, a common limitation of human decision-makers is that they do not trust the agents appropriately, resulting in either under-trust or over-trust [5, 6, 7, 8]. To overcome this limitation, we will tap the power of interaction and analogies [9, 10] to design personalized user training [11] to improve humans' AI literacy [12], helping them better understand the uncertainty inherent in the agents' learning and reasoning processes as well as the agents' limitations [13]. We will facilitate humans' development of an accurate mental model of the agents by studying (1) in the initial interaction stage, how to effectively communicate agents' behavior (e.g., error boundaries [14], rationale [15]) to humans using a mixture of deductive and inductive approaches; and (2) as the agents keep updating, how to enable humans to adjust their mental models accordingly by contrasting the agents' behavior with humans' existing mental models and justifying why these changes occur [16]. Finally, we will proactively predict whether humans' reliance on the agents will be appropriate, and provide just-in-time interventions by nudging humans to engage with the agents more cognitively [6], carefully framing the outcomes of different reliance decisions [17, 18], and encouraging humans to solicit a "third opinion" from other humans or agents.

2. **Modeling human behavior in agent-in-the-loop decision-making.** To quantify how human decision-makers react to agents, we will develop behavioral and cognitive models to characterize whether and how humans utilize assistance from agents during their decision-making. We will develop models to predict whether humans will follow the agents' decision recommendations, building upon theoretical frameworks in economics and cognitive science such as prospect theory [19, 20] and instance-based learning theory [21], while taking into account how the adoption of the agents' recommendations is affected by humans' own evaluation of the task [22] and social influences [23]. Similarly, when agents provide indirect intelligence to support humans' decision-making, we aim to model whether and how humans will direct their attention to and act upon the intelligence, given various properties of the intelligence, such as its relevance, presentation formats, frequency, and timing.

Designing agents that are aware of and adaptive to human behavior. Informed by the behavior and cognitive models identified above, we will re-design intelligent agents to

take humans' reactions to agents into account. For example, we will incorporate humans' behavior of adopting the agents' decision recommendations or not into the objective functions of the agents (e.g., loss function) to directly optimize the joint human-agent decision-making performance [24]. The knowledge of how humans will respond to intelligence supplied by the agents will also guide agents to adjust the ways in which they prioritize and group intelligence, interrupt at appropriate times and places, and suggest actionable steps that humans can take upon receiving the intelligence. We will use model tracing and dynamic parameter fitting [25, 26, 27] to continuously update the human models so that the agents can be adaptive to changes in human behavior and personalized for each individual human partner.

Thread 2: Learning and Reasoning with Humans-in-the-Loop. This paradigm reflects the scenarios in which human intelligence is solicited from a crowd of mixed expertise to augment the intelligence of autonomous agents for performing security-related tasks, such as the open-world threat knowledge extraction in Thrust AI-1. Traditionally, human intelligence has mainly been solicited through low-level, labeled data [28, 29]. We are interested in exploring, beyond labels, what additional high-level intelligence can be solicited from humans and how to effectively solicit it. We highlight two directions to this end.

First, we will investigate methods to solicit domain knowledge from humans. Humans' domain knowledge can come in various forms, such as causal relationships between different events, and identification of meaningful composite variables and missing variables. We will leverage BLOG [30] to enable humans to convey their domain knowledge to agents. To further facilitate such communication, we will design approaches for automatically translating humans' natural language description or graphical specification of their knowledge into BLOG code.

Second, we will investigate methods to solicit feedback from humans. We are particularly interested in feedback that reflects humans' intuition, such as their judgment on whether an action of an agent is "good" or not and whether the agents' behavior is understandable to them [31]. We will examine how instance-level feedback can be solicited in the most efficient manner, and we will also explore the potential of empowering humans to generate global, abstract feedback. Once the feedback is obtained, we will examine how it can be effectively used to guide the training of the agents (e.g., via influencing the reward function of a reinforcement-learning-based agent [32, 33]).

We note that for both humans' domain knowledge and their feedback, how to solicit them in a scalable and reliable manner is a critical challenge. To improve the scalability of human intelligence solicitation, we will investigate the potential of implicitly soliciting humans' domain knowledge and intuition from human behavioral data, such as the data reflecting how humans themselves take actions [34, 35]. Furthermore, we will examine how machine assistance can be utilized to "speed up" the human intelligence solicitation process. For soliciting domain knowledge, one use of machine assistance could be conducting automated analysis of relevant data, and then presenting those results that may reflect domain knowledge (e.g., candidate causal graphs obtained by applying structure learning algorithms on event logs [36, 37]) to humans for validation and improvement.

Thread 3: Human-Agent Teaming. This paradigm reflects the scenarios in which humans and agents are working on separate sets of tasks toward a common goal. We consider two types of human-agent teaming: division of labor and collaboration with interdependence.

Division of labor reflects the case where humans and agents simply share the workload, but the action/performance of one party has limited impacts on the other party. We will address this challenge by (1) taking a discriminative approach to learning a task allocation model considering the utility brought up by each agent on the task (e.g., task performance), or (2) by formulating this problem as a contextual bandit problem [38] to learn the optimal task allocation policy in an online fashion from scratch, or following the batch learning from bandit feedback framework [39] to learn the policy from historic performance logs of different agents completing various tasks.

Human-agent collaboration with interdependence reflects the case where humans and agents each have their own tasks to complete in parallel, and the action/performance of one party will affect the other party. The key challenge is to design effective communication strategies between humans and agents such that one party can see why the other party takes certain actions, be aware of environmental changes that will affect their own task completion, and understand the other party's expectations of them. To this end, we will extend the human-aware planning framework [40, 41] by modeling not only the agents' task model MA and humans' mental model of the agents MAh, but also the humans' task model MH and the agents' mental model of the humans MHa. The communication from agent a to human h can then include information "push" [42] to help humans update MAh, update MH given MHa, and provide action suggestions based on MHa. Moreover, agents can also "pull" information from humans, such as requesting updates from humans to update MA and MHa.

3.2.2 Planned Research Activities

To meet the challenges of Thrust AI-2, we will undertake the following projects during Year 4:

Project: Social Influence In Networked AI And Human Agents: Experiments And Theory

Leads: Jie Gao (Rutgers), Ming Yin (Purdue)

Abstract: AI agents increasingly operate in networked settings where they share information and jointly assess threats, yet little is known about how social influence shapes their behavior. Decades of behavioral research show that human decision-making in such settings is often distorted by conformity, peer pressure, and authority effects rather than being purely truth-seeking. Because LLM-based agents are trained on human-generated data and interact through social communication protocols, they may inherit similar social influence dynamics, with important implications for the reliability of multi-agent intelligence systems. This project centers on the question of when and why networked AI and human agents exhibit social influence effects, how such effects can be formally characterized, and how risks arising from social influence can be mitigated through principled design, particularly in high-stakes threat-intelligence settings.

Outcomes: This project will yield controlled experimental evidence on social influence in networked human-AI systems and theoretical models that explain and predict these effects. We will disseminate our research findings in leading HCI and AI/ML venues (e.g., CHI, CSCW, NeurIPS, ICML, ICLR, AAMAS).

Primary Thrust: AI-2
Secondary Thrust(s): AI-3 and SEC-3.

Project: Human-AI Decision Making for Protecting Critical Infrastructure

Leads: Nick Feamster (Chicago), Ambuj Singh (UCSB), Ming Yin (Purdue)

Abstract: Many cybersecurity decision problems can be naturally formulated as optimization problems that reduce risk under operational constraints. We will conduct human-AI studies to understand how such optimization problems can be formulated

and solved. We focus on two complementary infrastructure protection scenarios: (1) network-level DDoS mitigation, where an operator must decide which traffic flows to rate-limit, reroute, or drop across a multi-site infrastructure while preserving service to legitimate users, and (2) network anomaly detection and incident response, where an operator must allocate limited monitoring and analysis resources across network segments to identify compromised hosts, detect novel attack traffic, and contain lateral movement. This task studies how LLM-based planners can elicit soft constraints through interaction by comparing various human preference solicitation strategies. Possible strategies include optimal-plan feedback, contrastive plan comparison, and sequential "what-if" variations. Key evaluation metrics will include the quality of the resulting solutions after fixed interaction budgets, interaction efficiency, and humans' cognitive load and trust towards the LLM-based planner.

Primary Thrust: AI-2: Human-Agent Interaction.

Secondary Thrust(s): AI-4: Strategic Gaming. SEC-2: Threat Detection. SEC-4: Response and Recovery.

3.3 Thrust AI-3: Multi-Agent Collaboration

Agents collaborate not only with humans but also among themselves. We propose to develop a multi-agent collaboration framework, enabled by research in the following aspects: Modeling of multi-agent networks, multi-agent reasoning, and multi-agent adversary-aware learning.

Challenges:
- Current multi-agent frameworks lack an analytical model to assess the aggregate capability of a network of agents.
- Current multi-agent reasoning approaches do not consider agent heterogeneity and asynchrony, or absence of knowledge.
- Current multi-agent learning approaches underestimate the existence of adversarial agents and conflicting views of the world among agents.

We believe that these challenges, as well as the proposed research to address them, are common and fundamental to multi-agent collaboration. Hence, we expect that our research agenda and priority in this thrust will be stable but adjustable for the duration of the institute.

3.3.1 Proposed Research for Years 4-5

Thread 1: Statistical Models in Multi-Agent Networks. We consider a distributed network of agents that jointly perform a security function (e.g., detection). The model will include node (agent) attributes, including agent-level characteristics such as agent-specific knowledge, specialties (sensing, actuation, and decision-making capabilities), and tasks. The nodes' knowledge will be probabilistic in nature, for example, the chance or belief of certain cyber-attacks (Thrust AI-1). Our model will also include attributes on edges or hyper-edges, capturing properties of interactions and relationships among agents such as trust, information exchange, or aggregation patterns. Both node features and edge attributes are dynamically evolving. We propose a co-evolving model where node features, especially local knowledge of each agent, are exchanged, subject to the dynamics of edge attributes. The exchange pattern and frequency are determined by the agents' common interests and mutual trust level. We expect this model to be well-suited to characterize distributed agents sharing knowledge and correlating their individual discoveries (e.g., for detecting a "pandemic" attack).

Our multi-agent network model is inspired by the contagion model: Agents share their knowledge and new discoveries with other agents in a network such that all agents in the network can provide timely responses to emerging phenomena of interest. We will develop a suite of protocols for agents to communicate with one another using their new discoveries/knowledge, such that new phenomena that are independently detected and confirmed by some agents in the network can be quickly recognized and acted upon in a timely manner. The work builds upon prior literature on contagion models and network consensus [43, 44, 45, 46, 47, 48], but expects to embrace much more complex agent behaviors and interaction patterns.

Thread 2: Resilient Multi-Agent Reasoning. To enable reasoning in a multi-agent network, agents need to know which other agents to query for certain kinds of knowledge. We will address this problem by leveraging our prior work in distributed trust management, which involves a knowledge type system and the corresponding efficient discovery algorithm [49]. Here, each agent holds a local context that defines the knowledge it maintains (e.g., beliefs, environmental information, and trends, as proposed in Thrust AI-1), and identifies types of missing knowledge and from which agents the knowledge may be obtained. We will explore new approaches that combine probabilistic and quantifiable approaches with logical approaches such as defeasible speculative reasoning in a distributed agent setting [50]. We will combine the probabilistic programming language approach with approaches that are based on

answer-set Prolog, e.g., P-log [51], which combines defeasible logic with Bayesian probability nets and supports belief revision by means of Bayesian conditioning.

We have been focusing on how to aggregate information with uncertainty from agents in a network to improve the collective prediction accuracy in a multi-agent system. One pitfall we need to avoid is the well-known issue of information cascade (also called herding) in sequential learning, where wrong predictions due to bad luck early in the sequence can trigger all subsequent agents making wrong predictions even with fully rational agents. We have studied what types of network topologies and decision-making parameters can enable asymptotic truth learning in a binary setting.

Thread 3: Reinforcement Learning for Multi-Agent Systems. As agents operate in adversarial environments, they will need to be resilient to adversarial attacks and prevent an adversary from inferring information about system behavior. To synthesize resilient policies against adversarial actions, agents need to have a mechanism to learn about adversarial engagement.

Recent results indicate that a lack of suitable adversary-resilient training can result in multi-agent networks that are fragile to even the slightest adversarial disturbances [52, 53, 54]. Using MARL, we will investigate how the presence of adversaries will change trajectories and learned policies. We will use a clean dataset and MARL training for reward redistribution [55] and know the efficient trajectories in the absence of attacks. Then we will use adversary-embedded dataset (with our recent approach called FRESH [32]) for the same task, which consists of attacks selected from MITRE ATT&CK vectors [56] in conjunction with randomly injected adversarial (negative) reward inputs along the length of the trajectory to represent adversarial actions.

Also on reinforcement learning, we consider applications in robotics. In particular, Diffusion Policies have quickly gained prominence in Imitation Learning of Robotic control policies. Diffusion Policies have a unique ability to represent multimodal distributions, which are often necessary to model human demonstrations. In this project we seek to investigate the applicability of Diffusion Policies in reinforcement learning.

Thread 4: A Secure and Private Framework for LLM-powered AI Agents. We expect that autonomous agents powered by LLMs and Retrieval-Augmented Generation (RAG) will have increasing importance, both in cyber threat intelligence and operation and in other applications. We propose to develop a framework that regulates the interaction between agents and their interaction with local information to address malicious threats. A framework to help secure a local LLM-based AI agent could

include several components: limiting the access of the agent to critical and/or private local data so that they are used for the right context, filtering communications from external agents to detect and/or remove malicious contents, filtering outgoing communications to avoid leakage of sensitive/private data that is not needed for the task, restarting component agents to fresh, clean states to avoid accumulated impacts of past communications, and so on.

3.3.2 Planned Research Activities

In Year 4, we will undertake the following planned research projects to address the challenges and research threads outlined in Thrust AI-3.

Project: Auditing Agentic Unlearning: A Forensic Membership Inference Framework

Leads: Jie Gao (Rutgers) and Ninghui Li (Purdue)
Problem Statement: As AI agents continuously update their internal states through interaction, it is necessary to develop mechanisms to selectively remove specific knowledge subsets, such as toxic content or private data, to comply with safety and privacy regulations. This project seeks to formalize a forensic auditing framework that leverages membership inference to quantify residual statistical traces.

Proposed Research: We propose to solve the problem of verifying whether an AI agent has truly forgotten specific information it was instructed to "unlearn." To address this, we propose a forensic auditing framework that treats unlearning like a game of "find the target":
Primary Thrust: AI-3: Multi-Agent Collaboration.

Secondary Thrusts: SEC-2: Threat Detection and SEC-4: Response and Recovery. AI-1: Continual learning with safe recovery guarantees.

Project: Statistically Verified Representation Learning for Security-Critical Decision Systems

Leads: Xiaoming Huo (Georgia Tech) and Joao Hespanha (UCSB)

Abstract: Modern security settings rely on transforming high-dimensional data X into lower-dimensional representations Z to facilitate rapid decision-making. This project

proposes the Sufficient Representation Learning (SRL) framework to address the "Dilemma" of objectives: maximizing Efficiency (learning with finite samples) and maintaining Sufficiency (lossless information retention); for agents engaged in monitoring cybersecurity and physical sensors to actively respond to attacks.

Primary Thrust: AI-3: Multi-Agent Collaboration.
Secondary Thrust(s): AI-4 and SEC-4, focusing on moving beyond heuristics to provide certifiable statistical guarantees for learned representations in critical decision-making systems.

3.4 Thrust AI-4: Strategic Gaming and Tactical Planning

In coordination with human teams, intelligent security agents will need to make decisions to thwart attempted attacks and remove adversary agents from the system. These decisions must be made under uncertainty in an ever-changing cyber environment. We propose to establish a defensive, composable layer that realizes the agents' strategies and tactics, enabled by research in the following aspects: foundation of strategy games, adversary/threat modeling and defense function certification, and dynamic planning of agent tactics.

Challenges:
- Current game-theoretical approaches mainly focus on defending against anticipated attacks, with little reasoning about the dynamic nature of the attackers' behavior.
- 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.
- Current detection systems are incapable of dealing with fast and high-volume information flows.

We believe that this thrust represents a unique inspiration for foundational AI by cybersecurity, given the "arms race" and mutually-untrusted relation between adversaries and defenders. As such, this thrust will not only generate methods and techniques that are directly applicable to cybersecurity, but also lead to new research problems and concepts that make AI more strategically robust in "combative" application scenarios.

3.4.1 Proposed Research for Years 4-5

Thread 1: Foundation of Strategy Games. The interaction between the intelligent security agents and the adversaries, both making use of AI, can be viewed as dynamic non-cooperative cyber-AI games in stochastic environments. The adversaries aim to advance towards their end goal while remaining undetected (stealthy), while the agents aim to detect the adversarial presence, infer the intent, locate and possibly isolate or remove it before further progress inside the system. A realistic modeling of this strategic interaction is based on the observation that adversaries introduce anomalous information and control flows as they compromise programs and exfiltrate data, which can be detected via dynamic provenance tracking [57, 58]. We will investigate and develop suitable information and control flow representations, including graphical, which will lead to actionable game models based on these representations.

Two classes of games to be developed are detection games and deception games. The solutions to the detection game will provide a list of locations (or nodes in a flow graph) to inspect, enabling prioritization by the operator. The solutions to the deception game will provide the amount of information needed to improve defense strategies at the next step as well as the amount of biased information that needs to be introduced by the cyber sensors to mislead future adversarial action.

We will investigate and develop the needed new class of non-equilibrium-seeking cyber-AI games that aims to compute best responses of players at each stage of the cyber kill-chain. We will make use of our prior modeling on offline, postmortem detection of APTs (including ransomware, reconnaissance, and multi-stage nation-state attacks) using log data [59, 60, 61, 62, 63]. We will explore and extend the concept of learning-based regret minimization [64].

Thread 2: Adversary Modeling. Assessing the robustness of detection and mitigation strategies involves developing cyber attack models, with the understanding that such models are inherently incomplete, carry large uncertainty, and are typically hard to optimize against [65]. In this thrust we explore alternative approaches to adversary modeling for the cybersecurity domain.

Monte-Carlo-based probabilistic graphical models capture conditional dependencies between attacker actions and contextual factors such as adversary skill level, exploit availability, patch adoption rates, network topology, and system configurations. Sampling from empirically grounded distributions, agents explore a diverse set of

attack paths and their outcomes to generate empirical risk metrics, including expected loss distributions, breach probabilities, and time-to-compromise estimates.

Competitive Markov Decision Processes (CMDPs) extend Markov models to cybersecurity domains, where we need to capture the interaction between a cyber defense system and an agent that tries to evade IDS detection. Under the CMDP framework, the attacker seeks to optimize a reward function that balances attack success with the need for stealth. This modeling approach leads to a hybrid game formulation: an attacker-modeled CMDP nested within a Stackelberg game.

Thread 3: Dynamic Planning of Agent Tactics. Information asymmetry, which is inherent in cyber games, arises due to incomplete information about the adversary's capabilities as well as partial observability of the cyber "theater" by both security agents and adversaries and increases the complexity of solving the games. On the other hand, information asymmetry can be exploited by a variety of strategies that inform the defense, including deception [66], moving target defense [67], and information and network structure obfuscation [68, 69].

Developing precise cognitive models and maintaining deception over multiple rounds of interaction are open problems in deception-based defense [70] that we will explore. To quantify the worst-case and average-case performance of the deception tactics to be developed, we will explore lifelong learning techniques [71, 72] integrated with risk assessment. We will investigate and develop distributionally-robust algorithms to compute tactical plans that achieve guaranteed performance even when the belief state of the system has been biased by adversarial misinformation [73].

Another tactical goal in planning is the obfuscation of defense against adversarial probing. We will investigate and develop methodologies that will enhance resilience to probing by incorporating concepts from differential privacy as a design constraint during the agent training [69]. As recent advances in composition theorems [74, 75] demonstrate that robustness to adversarial corruption and differential privacy are synergistic, we will develop new training/planning mechanisms that are privacy-preserving by design and provide quantifiable robustness against query-based attacks.

3.4.2 Planned Research Activities

To meet the challenges of Thrust AI-4, we will undertake the following project during Year 4:

Project: Red Agent for Stealth Strategy Optimization in Blue Agent-Protected Industrial Control Systems

Leads: Radha Poovendran (University of Washington) and Dongyan Xu (Purdue University)

Abstract: The integration of collaborative "blue agents" into Industrial Control Systems (ICS) raises a fundamental question: do synchronized defenses introduce exploitable structural vulnerabilities? While blue agents share intelligence to harden defenses, an adaptive red agent can learn from past defeats to survive. Existing collaborative defense frameworks do not model attackers that reason about inter-target synchronization and adaptation. In this research, we will study whether a red agent that anticipates coordinated blue agents' behaviors can systematically degrade the latter's collective defense effectiveness. By exposing and modeling such vulnerabilities under a game-theoretic framework, we expect to reduce predictability and hence improve collective robustness of collaborative blue agents.

Motivation: In ICS such as the chemical and water-processing plants in the Year 4 scenario (Section 4), homogeneous IT architectures create a race of adaptation between attackers and defenders. Detection at one facility triggers synchronized rule updates across others. While this strengthens collective defense, it also produces predictable adaptation signals that a strategic adversary can observe and exploit.

Expected Results and Demo: For validation and evaluation, we will set up an experimental environment (on GATE) featuring two simulated blue agents operating in distinct industrial facilities interconnected by an intelligence-sharing channel. We will quantitatively evaluate (1) reduction in detection probability following synchronization events, (2) increase in adversarial persistence time across facilities, (3) accuracy of red-agent prediction of synchronized detection rules, and (4) the transferability of adaptive policies between heterogeneous ICS environments.

Primary Thrust: AI-4: Strategic Gaming.

Secondary Thrust(s): AI-1: Learning and Reasoning. AI-3: Multi-Agent Collaboration. SEC-2: Threat Detection. SEC-4: Response and Recovery.

Project Title	Collaborators	Primary Thrust(s)	Secondary Thrust(s)
Self-Motivated and Continuous Learning Agents for Security Applications	UI Chicago & Georgia Tech	AI-1	AI-2, SEC-2
Assessing and Mitigating Inherited Vulnerabilities in Multi-Modal Foundation Models for Agentic Defense of Critical Infrastructure	Purdue & U Washington	AI-1	AI-3, AI-4, SEC-1, SEC-2, SEC-3, SEC-4
Secure Agents for Cyber and Patient Interactions	UC Berkeley & UI Urbana-Champaign	AI-1	AI-3, SEC-1
Latent-Space Detection Systems	UC Santa Barbara & U Virginia	AI-1	SEC-2, SEC-4
Social Influence In Networked AI And Human Agents: Experiments And Theory	Rutgers & Purdue	AI-2	AI-3, SEC-3
Human-AI Decision Making for Protecting Critical Infrastructure	U Chicago, UC Santa Barbara & Purdue	AI-2	AI-4, SEC-2, SEC-4
Auditing Agentic Unlearning: A Forensic Membership Inference Framework	Rutgers & Purdue	AI-3	AI-1, SEC-2, SEC-4
Statistically Verified Representation Learning for Security-Critical Decision Systems	Georgia Tech & UC Santa Barbara	AI-3	AI-4, SEC-4
Red Agent for Stealth Strategy Optimization in Blue Agent-Protected Industrial Control Systems	U Washington & Purdue	AI-4	AI-1, AI-3, SEC-2, SEC-4

Table 1: Year 4 ACTION Foundational AI Research Projects

3.5 Timeline of Activities

The projects described above as "Planned Research Activities" are scheduled to occur between June 2026 and May 2027. All projects were selected by a review panel consisting of the ACTION Executive Committee and the eight thrust leads after comparison between project proposals and this S&IP. Projects are required to be

collaborative, led by at least two faculty researchers from separate universities. Project leads are required to report on progress during quarterly and annual reporting.

4 Use-Inspired Research

The ACTION Institute's use-inspired research will continue to be driven by four main thrusts, which correspond to the four main steps of the cyber-defense life-cycle: assess, detect, attribute, and respond/recover. The research under these thrusts builds on the AI stack for agents, enabled by the foundational AI research thrusts, to realize novel security capabilities of intelligent agents. In return, it provides domain usage scenarios, requirements, and challenges to the AI thrusts to motivate new foundational research.

Our Year 4 research activities are inspired by and connected through a narrowed version of the overarching scenario presented in Section 2.1. To focus our efforts, we specifically consider a situation in which a nation-state actor is conducting a multi-pronged attack against the United States' critical infrastructure.

The operation's goal is to create uncertainty and chaos by compromising two groups of industrial plants, each with a (somewhat) homogeneous IT infrastructure: chemical plants; and water-processing plants.

The similarity of IT infrastructure allows the attacker to reuse their procedures when attacking instances within the same group (e.g., attacks on chemical plants follow similar steps, whereas attacks on water-processing plants follow different steps).

These plants have intelligent security agents for protection. The agents rely on both local and global infrastructure to store, process, and share knowledge. They work together with the goals of:

- Continuously assess the security posture of the infrastructure (SEC-1);
- Identify attacks on networks, endpoints, and services (SEC-2);
- Produce, share, and consume threat intelligence about attack instances (SEC-3);
- Respond to attacks and recover from them (SEC-4).

The agents are supported by a series of AI capabilities that allow them to:

- Continuously learn from their environment and reason about the steps to take (AI-1);
- Interact with the operators of the plants' infrastructure to capture priorities and directives (AI-2);

- Collaborate to support the security process in novel, unexpected ways (AI-3);
- Devise dynamic, reasoning-based collaborative plans that reflect the current status of the infrastructure (AI-4).

Beyond the scenario, ACTION faculty collectively presented a list of supporting capabilities that could result from research. Those capabilities led to the 10 planned research activities to be pursued during the next reporting period. Of note, we associate each of the activities with a single ACTION SEC thrust. We also identify secondary and supporting thrusts for each research activity, with the intent of demonstrating interconnectedness.

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

Proactively identifying vulnerabilities and assessing their impacts in complex, large-scale systems is of paramount importance to reduce the attack surface and/or prevent cyber exploitations. The increasing complexity and dynamics of modern software and hardware (including the AI components), present a significant challenge for vulnerability discovery.

Challenges:
- Traditional static and dynamic analysis techniques struggle to scale effectively.
- Static methods often generate excessive false positives and require extensive manual triage.
- Dynamic approaches suffer from limited coverage and execution constraints.

To address these challenges, we propose a novel AI-agent-powered framework that integrates traditional vulnerability analysis with novel machine learning and large language model (LLM)-based techniques. Our goal is to significantly improve the efficiency and effectiveness of vulnerability discovery, understanding, and remediation. We will focus on both (1) improving the vulnerability assessment of individual system components (e.g., software/hardware systems, AI modules) and (2) performing holistic vulnerability assessment of composite systems (e.g., a critical infrastructure).

4.1.1 Proposed Research for Years 4-5

Thread 1: Vulnerability Assessment of Complex Software using AI Agents. AI agents and LLMs offer new possibilities to significantly improve the efficiency and

effectiveness of vulnerability assessment for complex software. One direction we want to pursue is to use AI agents to generate "heatmaps" to highlight high-risk regions within large software projects. The second direction is to leverage AI agents to augment traditional methods such as fuzzing and symbolic execution. The AI agent can serve as an orchestrator (e.g., dynamically determining which techniques to apply in different contexts) or act as one of the executors (e.g., directly working with a fuzzer to increase analysis coverage). The third direction is to enable effective vulnerability scanning within dynamic environments through a Dynamic Profiling Agent (DPA) that leverages Monte Carlo simulation to focus localized validation and deep testing on the most vulnerable points.

Thread 2: Patching Software Vulnerabilities with AI-Driven Code Reasoning. Once vulnerabilities are identified, addressing them efficiently is crucial. A promising direction is to develop an AI-powered agentic patching system that leverages LLMs to analyze code dependencies, reason about vulnerability sources, and suggest security fixes. Unlike naive code generation approaches, patch generation will need to incorporate structured exploration techniques to traverse the source code tree, ensuring that patches do not introduce regressions or new vulnerabilities.

Thread 3: Vulnerability Assessment of AI Agents Themselves. In addition to using AI agents and LLMs for vulnerability assessment, we also want to gain a deeper understanding of the vulnerabilities in AI agents and LLMs themselves. First, AI agents and LLMs are the critical building blocks for the proposed defense system, and thus their security should be carefully vetted. Second, users today are actively using LLMs for real-world software development. If the LLM-generated code contains vulnerabilities, it could be exploited by malicious parties. To this end, we want to understand the effectiveness of potential attacks against LLMs and agents, such as prompt injections and malicious fine-tuning, and explore effective countermeasures.

Thread 4: System-wide Vulnerability Assessment of Cyber-Physical Systems. We plan to extend the analysis from individual components to the entire system for a system-wide vulnerability assessment. More concretely, we will focus on cyber-physical systems, such as critical infrastructures. Our vision is to use the learning and reasoning capabilities of AI agents to enable new ways of vulnerability discovery and assessment for CIs. We envision agents that specialize in CI modeling, simulation (digital twining), and vulnerability discovery. We will also develop a Cyber-Physical Knowledge Graph (CPKG) where the graph edges explicitly link physical and cyber nodes to represent the interactions and influences between them.

4.1.2 Planned Research Activities

To meet the challenges of Thrust SEC-1, we will undertake the following projects during Year 4:

Project: Exploring Tokenizer-based Attack Surfaces for LLMs

Leads: Ninghui Li (Purdue) and Sewoong Oh (U Washington)

Abstract: The proposed research addresses fundamental properties of LLMs which will be used by the agents in the overarching scenario. As Large Language Models (LLMs) are trained on internet-scale data with little supervision or curation, membership inference attacks (MIAs) have emerged as an important tool to quantify the extent to which a model leaks information. In this project, we propose exploring a new and highly promising attack vector: the LLM tokenizer. A tokenizer is a fundamental component that serves as the gateway to an LLM, converting raw text into discrete tokens. Tokenizers are uniquely suited for membership inference for three key reasons: (1) A tokenizer's vocabulary is learned directly from a text collection representative of the model's overall pre-training corpus. (2) Tokenizers are universally open-sourced, even for closed models such as OpenAI's GPT series. (3) Unlike LLMs, tokenizers are significantly smaller and can be efficiently trained from scratch.

Primary Thrust: SEC-1
Secondary Thrust(s): AI-1. SEC-2.

Project: A Dataset for Evaluating AI-Based Logic-Vulnerability Analysis

Leads: Giovanni Vigna (UC Santa Barbara) and Tonya Fields (Norfolk State University)

Abstract: Large language models (LLMs) are a promising direction for building automated tools that can reason about higher-level intent, infer implicit requirements from code and documentation, and surface security-relevant deviations. However, progress in this area is currently bottlenecked by a lack of datasets that directly support the evaluation of systems designed to identify logic vulnerabilities. We propose to develop a novel dataset for evaluating automated systems that identify logic vulnerabilities. Our dataset will consist of small-to-medium applications paired with (i) an explicit description of intended behavior, (ii) one or more injected logic

vulnerabilities grounded in that intended behavior, and (iii) metadata that supports rigorous evaluation.

Primary Thrust: SEC-1 (Threat Assessment)

Secondary Thrust(s): AI-1 (Reasoning) and AI-2: Human-Agent Interaction.

Project: Assessing and Mitigating Cognitive Biases in LLM-Based Vulnerability Detection

Leads: Z. Berkay Celik (Purdue) and Gang Wang (University of Illinois Urbana-Champaign)

Project Description: This project investigates the presence of human-like cognitive biases, specifically the Halo Effect, the Framing Effect, and the Anchoring Effect, in Large Language Models (LLMs) used for vulnerability detection. We will quantify these biases in the context of vulnerability detection and leverage the lessons learned to construct agents that prioritize semantic correctness over irrelevant contextual signals.

GATE Experiments: We plan to leverage the GATE environment to simulate realistic software development scenarios including misleading contextual cues, such as commit messages from senior architects, to test agent resilience.

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

Secondary Thrust(s): AI-1 (Learning and Reasoning with Domain Knowledge)

Project: Generative Causal Modeling for Multistage Attack Detection

Leads: Wenke Lee (Georgia Institute of Technology), Radha Poovendran (University of Washington), Nick Feamster (University of Chicago), Z. Berkay Celik (Purdue University)

Project Description: This project proposes a generative framework to address the scarcity of training data for multistage attacks in complex networked system environments. We posit that by integrating Diffusion Models with Neurosymbolic State-Space Models, we can generate new, high-fidelity, and continuous attack scenarios

that preserve the logical invariants (e.g., causality between privilege escalation and data exfiltration). We will use these synthetic environments to train defense agents in a game-theoretic simulation, which will enable them to detect novel attack patterns across diverse domains.

GATE Experiments: We will leverage the GATE environment to simulate the research scenarios. Specifically, we plan to deploy the Neurosymbolic Generator Agent within GATE to produce realistic, multistage attack traffic.

Primary Thrust: SEC-1

Secondary Thrust(s): SEC-2, SEC-3, AI-1, and AI-4.

4.1.3 Metrics and Outcomes

The success of the proposed research activities will be evaluated in the following ways. First, we will measure the size and diversity of the collected, curated, and released datasets of vulnerabilities and threat intelligence from this project. Second, we will evaluate the performance of vulnerability detection methods based on metrics including precision, recall, and computational costs. Third, we will evaluate the transferability and generalizability of the proposed system over time and across system domains.

The success will be also quantified by the number of newly discovered vulnerabilities (i.e., 0-day vulnerabilities) from this project. The integration with the AI stack will be focused on the type and the amount of modification needed to customize the generic AI models for practical use in vulnerability assessment tasks. In addition, we will evaluate how much the vulnerability assessment results can improve the efficacy of other downstream tasks in Thrusts SEC-2, 3, and 4 for threat detection, attribution, and response.

4.2 Thrust SEC-2: Intelligent Agents for Detection and Understanding

Detection of complex, multi-step attacks requires monitoring disparate telemetry and correlating security events across multiple domains and abstraction levels [76]. Once an attack is detected, it needs to be analyzed and described in a way that is useful to both machines and human analysts.

We propose to develop approaches that provide: robust training of detection agents, handling of threat evolution and evasion, and novel detection approaches incorporating uncertainty measures and multi-agent collaboration.

Challenges:
- Current learning approaches are brittle and degrade when facing incomplete, low-fidelity training data. Sharing data across organizations may be necessary, but privacy and security concerns must be addressed.
- Current AI approaches are inadequate in the cybersecurity domain due to adaptive and evasive adversaries.
- Current detection methods are unreliable and result in low accuracy in uncertain contexts, and produce models that are difficult to use and interpret.

Detection is a broad and well-established research topic in cybersecurity, with a scope larger than what we propose in this thrust. However, we believe that the threads we propose can address important limitations of current approaches and can benefit from our foundational AI research.

4.2.1 Proposed Research for Years 4-5

Thread 1: Robust Training of Detection Agents for Changing Environment. We will continue work on robust training of AI agents that can detect and understand threats. This includes handling concept drift and using approaches based on transfer learning to develop more robust detection of malware and malicious activity with less data, building upon our work on adapting to changes in Windows malware using transfer learning [77]. We will research methods to enhance robustness against increasingly sophisticated model evasion attacks.

We will also explore opportunities for using advances in Prediction-Powered Inference (PPI) [78] techniques for adapting data from a source domain to run inference on a target domain. We will also develop a controllable LoRA Soup that we can adjust in real-time as attacks and environments change. Given the exciting advances in language models and their reasoning capabilities, our research in this direction will involve developing methods for tuning an LLM that can analyze potential attacks or real-time data and determine when attacks have occurred.

Thread 2: AI-Enabled Threat Reproduction. Effective threat detection models depend on the availability of high-quality data, which necessitates the ability to reproduce or synthesize real-world threats. While CVEs provide essential information about

vulnerabilities, they lack the necessary context and infrastructure for practical reproduction. Addressing these challenges requires a paradigm shift from manual, ad-hoc reproduction efforts to an AI-driven, automated threat reproduction framework that can systematically generate attack scenarios.

Thread 3: Supporting Data Sharing. Although developing better threat detection and understanding systems benefits from aggregating data across many organizations, collecting datasets from multiple sources raises privacy concerns as well as security risks including data poisoning attacks. We want to develop methods for learning across organizations that both (i) preserve privacy for all participants in the collective dataset, and (ii) are robust against data poisoning attacks.

Thread 4: Multi-Agent Detection Systems. Traditional intrusion detection systems (IDS) tend to operate in isolation and struggle to detect sophisticated, multi-layered attacks that span across an organization's infrastructure. To overcome these limitations, we proposed a framework that deploys a network of heterogeneous Detection Agents that leverage multi-agent collaboration (AI Thrust 3) for anomaly detection and deception. We will also perform an in-depth analysis of the learning and reasoning methods and their applicability to stealthy behavior detection. We plan to work closely with core AI colleagues to improve and adapt the ML methods for more effective APT detection and threat hunting.

4.2.2 Planned Research Activities

To meet the challenges of Thrust SEC-2, we will undertake the following projects during Year 4:

Project: Agent Sandbox: Agentic Red Teaming Platform for Multi-Agent Systems at Scale

Leads: Dawn Song (UC Berkeley) and Bo Li (University of Illinois Urbana-Champaign)

Project Description: Autonomous AI agents are rapidly becoming a foundational abstraction for modern AI systems. A fundamental limitation today is the lack of a controlled yet realistic sandbox for agents. In this project, we aim to develop a comprehensive agentic red teaming platform as a sandbox for agents red teaming and risk assessment.

Outcomes: This sandbox will allow researchers and practitioners to safely, systematically, and repeatably test agentic and multi-agent systems under realistic operational conditions.

Primary Thrust: SEC-2: Threat Detection.

Secondary Thrusts: AI-1: Learning and Reasoning. AI-2: Human-Agent Interaction. AI-3: Multi-Agent Collaboration. SEC-1: Threat Assessment. SEC-3: Threat Intelligence/Attribution.

Project: Improve AI-Powered Cybersecurity through Deepfake Detection

Leads: Yung-Hsiang Lu (Purdue University) and Christopher Kruegel (UC Santa Barbara)

Abstract: This project will study deepfake detection methods motivated by how deepfakes could be used to influence security operations. Our technical focus is system-aware acceleration of deepfake detection to enable efficient deepfake detection across servers, laptops, and mobile phones.

Description: Computer-generated media (deepfakes), including images, videos, and audio, is now prevalent on the Internet. Deepfakes threaten cybersecurity by impersonating key personnel, colleagues, or family members to obtain unauthorized access to critical systems. This project will study deepfake detection methods focusing on how to make deepfake detection efficient across servers, laptops, and mobile phones.

Primary Thrust: SEC-2: Threat Detection

Secondary Thrusts: SEC-1: Threat Assessment. AI-3: Multi-Agent Collaboration. AI-2: Human-Agent Interaction.

4.2.3 Metrics and Integration

The success of the proposed research activities will be evaluated in the following ways. First, we will evaluate the accuracy of agents to detect attacks, using well-established metrics such as false negatives and false positives. Second, we will evaluate accuracy over longitudinal data. Third, we will conduct user studies to assess the ability of agents

to explain attacks. Finally, we will assess the cost (time, memory) of training agents and their inference time.

4.2.4 Reference Applications and GATE Testing

Using GATE and the overarching smart city example, we will showcase the capabilities of our intelligent agents for proactive security threat detection. These agents will operate at two levels: (a) identifying threats within individual system components (e.g., malware on a water treatment plant workstation) and (b) detecting multi-stage breaches spanning across the entire system (e.g., a coordinated attack compromising the smart grid and then pivoting to manipulate control commands).

Through human-in-the-loop testing within GATE, we can simulate evolving attacker behaviors and assess the agents' ability to adapt and maintain high detection accuracy.

4.3 Thrust SEC-3: AI-Driven Cyber Threat Intelligence (CTI)

Cyber Threat Intelligence (CTI) plays a central role in modern cybersecurity operations, enabling proactive detection and mitigation of threats. We propose an AI-driven system that minimizes the various challenges to identifying, analyzing, and responding to cyber threats in real-time.

Challenges:
- Extracting intelligence from raw logs involves significant manual effort and is error-prone.
- Sharing CTI among organizations is often ineffective due to disparate formats.
- Due to their manual workflows, CTI applications are often slow, and unable to operate in real-time.
- Current AI techniques are unable to classify malware corpora with high precision and accuracy.

4.3.1 Proposed Research for Years 4-5

Thread 1: AI-Driven Extraction of Cyber Threat Intelligence from Raw Attack Logs. Traditionally, extracting intelligence from raw logs involves significant manual effort, requiring analysts to parse logs, identify indicators of compromise (IoCs), and correlate them with known attack patterns. We propose leveraging large language models (LLMs) and deep learning techniques to automate the extraction of CTI from logs. Our

prior work on synthetic network traffic generation, NetDiffusion [79] and NetSSM [80], provides a foundation for addressing data scarcity in the CTI extraction context. Our work on end-to-end optimization of ML-based traffic analysis pipelines (CATO) [81] provides techniques for identifying causal relationships among events in heterogeneous logs.

Thread 2: AI-Powered Codification and Sharing of Cyber Threat Intelligence. The current process of encoding CTI into structured formats (e.g., STIX) is largely manual and error-prone, making real-time sharing difficult. We propose an AI-driven codification pipeline that includes: Automated CTI Structuring using NLP models, Context-Aware Sharing leveraging AI to enhance CTI with metadata, and Federated Learning for CTI Exchange developing decentralized AI models that learn from multiple organizations while preserving data confidentiality.

Thread 3: AI-Assisted Application of Cyber Threat Intelligence to Network Security Posture. Current CTI application workflows are largely manual, limiting their speed and effectiveness. We propose an AI-driven system that automates CTI application by: Threat Hunting with AI, Automated Defense Adjustments implementing reinforcement learning-based models, and CTI Impact Analysis using AI to assess whether a given threat intelligence feed is relevant to an organization's specific network topology. Our generative models (NetDiffusion, NetSSM) enable controlled red-team-style evaluation of automated defense adjustments by synthesizing adversarial traffic conditions against which RL-based controllers can be stress-tested before deployment.

Thread 4: AI-Driven Malware Classification. While AI/ML techniques have shown promise in malware family classification, current approaches fall short of adequately classifying diverse malware corpora with high precision and accuracy. We plan to create a framework that automates malware binary processing to produce standardized datasets for AI/ML analytics and establishes a heterogeneous classification ensemble to enhance malware family attribution. In February 2025, we released our first malware analysis dataset for AI/ML modeling called MABEL (Malware Analysis Benchmark for AI and ML) to NAIRR.

4.3.2 Planned Research Activities

To meet the challenges of Thrust SEC-3, we will undertake the following projects during Year 4:

Project: Towards an Agentic Malware Intelligence Community

Leads: Giovanni Vigna (UC Santa Barbara) and Dongyan Xu (Purdue University)

Problem Statement: Malware proliferation continues to increase, posing significant threats to cyber and cyber-physical infrastructures, typically as part of APT attacks launched by well-resourced adversaries. Building on our Year 3 malware attribution research, we envision an agentic, cooperative malware defense paradigm that moves beyond individual classification and towards coordinated analysis, reasoning, and response. We propose to develop an LLM-based, multi-agent malware "intelligence community" that unifies detection, triaging/classification, and capability reasoning via the creation of a standardized malware capability taxonomy as well as a community of agents capable of cooperative malware analysis, classification, and response.

Proposed Work: Building on our Year 2–3 research, which produced and extended MABEL, one of the first public large-scale and AI-ready malware datasets, we propose to develop an LLM-augmented, multi-agent malware intelligence community, via three research tasks:
Outcomes: Our malware intelligence community will be evaluated using the MABEL corpus of malware samples attributed to respective families, as well as newly synthesized samples targeting the critical infrastructures in the overarching scenario.

Primary Thrust: SEC-3: Threat Intelligence/Attribution

Secondary Thrust(s): AI-1: Learning and Reasoning. AI-3: Multi-Agent Collaboration. SEC-2: Threat Detection. SEC-4: Response and Recovery.

Project: AI-Assisted Network Security Posture Management

Leads: Nick Feamster (University of Chicago) and Wenke Lee (Georgia Institute of Technology)

Problem Statement: Cyber Threat Intelligence is only as valuable as an organization's ability to apply it to live network telemetry. In practice, the gap between CTI feeds and operational defenses is substantial: feeds arrive in heterogeneous formats, are often stale or only loosely contextualized to a defender's specific topology, and the ML-based classifiers used to act on them degrade under distribution shift in production traffic. Building on our group's prior work on generative traffic models, end-to-end optimization of ML-based traffic analyzers, and robust traffic classification under shift,

we propose a research program that closes the loop between CTI feeds and operational defenses.

Outcomes: All artifacts will be evaluated within GATE on the smart-city overarching scenario, and will integrate with the malware intelligence community project (Towards an Agentic Malware Intelligence Community) to support end-to-end CTI-driven defense against coordinated APT attacks.

Primary Thrust: SEC-3: Threat Intelligence/Attribution.

Secondary Thrust(s): SEC-2: Threat Detection. AI-1: Learning and Reasoning. AI-3: Multi-Agent Collaboration.

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

Attack response and recovery is a critical component of defense-in-depth. We propose to leverage our foundational AI research to develop a response and recovery (R&R) approach that is based on cooperating intelligent security agents, capable of devising and executing R&R strategies and tactics [82]. Our approach is supported by the collaboration, learning, and reasoning primitives of our proposed AI stack. The R&R Agents will use the knowledge collected by the assessment, detection, and attribution agents (Thrusts SEC-1, SEC-2, and SEC-3) on both real-world and simulated worst-case attacks to set the optimal R&R plan, guided by the adversary models and corresponding defense function certificates (Thrust AI-4).

This thrust will develop novel approaches to response and recovery that focus on: Analysis and containment of multi-step attacks, robust and scalable forensics against manipulation attacks, and planning and adaptation of response and recovery.

Challenges:
- Current response approaches are not effective against dynamically evolving multi-step attack scenarios and new/rare events.
- Current approaches are not robust against sophisticated adversarial manipulations.
- Current response and recovery approaches are ad-hoc and specific to attack scenarios.

4.4.1 Proposed Research for Years 4-5

Thread 1: A System for Intelligently and Autonomously Managing Attacks. As modern cyber threats grow more sophisticated, defenders increasingly struggle to detect, analyze, and respond to attacks quickly enough to prevent serious damage. Existing security automation approaches, such as traditional Security Orchestration, Automation, and Response (SOAR) systems, rely heavily on pre-defined playbooks that lack flexibility when confronted with novel or evolving threats. At the core of the system is a reasoning engine that continuously analyzes the state of the system and the network, constructing an evolving model of the environment and any potential attacks in progress. We will base attack progression on well-known attack models, such as the Cyber-Kill Chain (CKC) and the MITRE ATT&CK. Our framework will integrate a suite of observation tools that allow agents to query hosts, network devices, and other infrastructure for relevant data. These tools will be selected dynamically by the agents themselves, based on their evolving understanding of the threat environment. Complementing these observation tools is a second class of tools dedicated to action and response, including mechanisms for host isolation, process termination, credential revocation, traffic filtering, and forensic data capture.

Thread 2: Incorporating Multiple, Collaborative Agents for Complete Situational Awareness. A critical aspect of this research will focus on how agents collaboratively reason about attacks that span multiple hosts or network segments. Each agent operates with localized visibility and control, but by sharing selected observations and response actions with peers, they collectively construct a distributed situational awareness. Through collaborative reasoning, agents can recognize subtle indicators across multiple systems that would be missed by isolated analysis, and they can coordinate response actions to block lateral movement or contain threats before they spread further.

Thread 3: Incorporating RL and GATE. The proposed system will incorporate a reinforcement learning framework in which agents are trained in GATE, repeatedly confronting realistic attack scenarios and variations of these attacks to learn effective observation and response strategies. We will create a large repository of attacks, classified based on possible attack goals, such as data theft, sabotage, denial of service, to enhance the training of agents in attack response and post-attack system hardening.

Thread 4: Hardening the Agents Against Attack. As the system relies on intelligent agents, a critical requirement is to also make them secure and reliable. The system will support methodologies for:

1. Building and evolving agents, particularly end-host agents and network agents with LLM interfaces, to make them secure against a landscape of evolving, stealthy attacks,

2. Ensuring the controllability and consistent behavior of LLM agents, if present, so one can specify guardrails and rules that will control their behavior during attack response and recovery, and
3. Automatically generating secure code (including patches) and executing system hardening operations (e.g., patching, compartmentalization, and camouflaging).

Building on Year 3 results in constraint-driven agent safety, including work on Secure ReAct and prompt injection defense via preference optimization, Year 4 will extend these foundations to address the full range of agent hardening requirements described above, with a particular focus on controllability guarantees and automated secure code generation.

Finally, the agents should be accountable and able to explain their decisions and actions.

4.4.2 Planned Research Activities

The research in Thrust SEC-4 is entwined in the efforts and success of all other thrusts. At least one planned project furthering SEC-2 also contributes significantly to all threads of SEC-4. Below, we describe the planned projects for Year 4 that address the challenges identified in this research thrust.

Project: Rollback Protocols for Agentic Memory for Resilient Response and Recovery in Critical Infrastructure

Leads: Elisa Bertino (Purdue), Bing Liu (UIC)

Motivation and Scenario Alignment: In the ACTION Year 4 scenario, intelligent security agents defend distributed chemical and water-processing plants while learning continuously from attacks, sharing threat intelligence, and coordinating responses. In such long-lived deployments, response and recovery (SEC-4) depend not only on restoring services, but also on restoring agent cognition after failures. When agents incorporate incorrect, poisoned, or misleading knowledge during an attack, those errors can persist and distort future responses unless explicitly undone. This project addresses a core security gap by developing rollback protocols for agent memory, enabling agents to recover semantically, not just operationally, from attacks, misalignment, or erroneous learning.

GATE / AI-Stack Integration: Rollback protocols will be exposed through standardized interfaces within the ACTION AI primitives provided by the AI-Stack, enabling recovery-aware reasoning and planning components to explicitly invoke memory rollback as part of response workflows.

Primary Thrust: SEC-4: Agent response and recovery through semantic rollback of compromised memory.

Secondary Thrust(s): SEC-2: Detection-triggered unlearning of incorrect or poisoned knowledge. AI-1: Lifelong learning with safe recovery guarantees. AI-3: Collaborative agents with coordinated recovery across sites.

Project: Agentic Vulnerability Discovery and Remediation

Leads: Wenke Lee (Georgia Tech), Dawn Song (UC Berkeley), Christopher Kruegel (UC Santa Barbara)

Abstract: In this project, we will focus on automatically identifying and remediating security bugs in real-world software. We propose novel reasoning approaches that improve the effectiveness and faithfulness of LLM-based agents, treating security as a first-class constraint rather than an afterthought in the software development lifecycle.

Outcomes: The final deliverable will be a system that can analyze large, real-world applications to identify a range of security flaws and propose remediations (e.g., patches) that eliminate root causes, with security constraints enforced throughout the agentic loop.

Primary Thrust: SEC-4: Response and Recovery.

Secondary Thrust(s): SEC-1: Threat Assessment and AI-3: Multi-Agent Collaboration.

Project Title	Collaborators	Primary Thrust(s)	Secondary Thrust(s)
Exploring Tokenizer-based Attack Surfaces for LLMs	Purdue & U Washington	SEC-1	AI-1 and SEC-2
A Dataset for Evaluating AI-Based	UC Santa Barbara &	SEC-1	AI-1, AI-2

Logic-Vulnerability Analysis	Norfolk State		
Assessing and Mitigating Cognitive Biases in LLM-Based Vulnerability Detection	Purdue & UI Urbana-Champaign	SEC-1	AI-1
Generative Causal Modeling for Multistage Attack Detection	Georgia Tech & U Washington & U Chicago & Purdue	SEC-1	SEC-2, SEC-3, AI-1, AI-4
Agent Sandbox: Agentic Red Teaming Platform for Multi-Agent Systems at Scale	UC Berkeley & UI Urbana-Champaign	SEC-2	AI-1, AI-2, AI-3, SEC-1, SEC-3
Improve AI-Powered Cybersecurity through Deepfake Detection	Purdue & UC Santa Barbara	SEC-2	SEC-1, AI-3, AI-2
Towards an Agentic Malware Intelligence Community	Purdue & UC Santa Barbara	SEC-3	AI-1, AI-3, SEC-2, SEC-4
AI-Assisted Network Security Posture Management	U Chicago & Georgia Tech	SEC-3	SEC-2, AI-1, AI-3
Rollback Protocols for Agentic Memory for Resilient Response and Recovery in Critical Infrastructure	Purdue & UI Chicago	SEC-4	SEC-2, AI-1, AI-3
Agentic Vulnerability Discovery and Remediation	Georgia Tech & UC Berkeley & UC Santa Barbara	SEC-4	SEC-1, AI-3

Table 2: Year 4 ACTION Use-Inspired projects

4.4.3 Metrics and Integration

A "correct" response and recovery can be defined as one that is applied to the correct target and achieves the intended outcome, e.g., patching the correct vulnerability so that it can no longer be exploited. Therefore, we will measure the accuracy of target identification and the effectiveness of action using situation-specific analyses (e.g., root-cause analysis, correlation, etc.) and verification and experimental evaluation (e.g., formal analysis, repeat/test, etc.). For attacks that span multiple networks and impact critical social services, a "correct" response and recovery can be one that provides the best available alternative rather than the default technical decision. We will evaluate our adaptive response strategy by formally verifying its decision-making/optimization algorithms and experimentally evaluating them using datasets of complex attacks.

4.5 Timeline of Activities

The use-inspired projects follow the same schedule, selection process and reporting requirements described in Section 3.5.

Table 2 above provides an overview of use-inspired research projects planned for Year 4 of ACTION.

5 Research Risks and Mitigation

There continue to be risks associated with the ACTION research and activities planned for Year 4 through Year 5 (2026-2028). Many of these risks were identified in earlier versions of this plan. We list them here as continued threats, as they remain persistent. We identify additional risks that relate to Education and Workforce Development and Broadening Participation below in the respective sections.

Risk 1: Lack of real-world, ACTION-ready data for research (continued risk). The proposed AI stack and its application to cybersecurity both require large amounts of real-world data from various application domains for ACTION knowledge extraction, training, and testing. Datasets for cybersecurity research, such as malware samples, program execution traces, and system operation logs, are more diverse and semantically complex to collect and curate than those in more established AI domains. The lack of "ACTION-ready" datasets may critically affect the development and evaluation of both our AI and cybersecurity research thrusts.

Mitigation: Year 3 produced several significant dataset contributions that directly address this concern, including TOUCAN, MABEL, CVE-GENIE, NetDiffusion, and NetSSM (see Sections 4.3 and 9). In Year 4, we will continue to identify cybersecurity datasets that are either publicly available or shareable from ACTION researchers or their collaborators, perform literature searches for datasets referenced in relevant papers, and reach out to authors to establish data sharing agreements. Such datasets will be uniformly formatted to support ACTION's AI and SEC research.

Risk 2: Lack of realistic, dedicated cyber/cyber-physical environments and shared high-performance AI processing infrastructures (continued risk). ACTION requires specialized infrastructures for experimentation and computation. For experimentation, a dedicated, realistic cyber/cyber-physical environment is desirable to implement and evaluate attacks and defenses. For computation, a shared, high-performance infrastructure is needed to support the data-intensive AI workload.

Mitigation: GATE has matured significantly since its initial deployment. Entering Year 4, GATE has a functioning architecture with defined interfaces, a growing library of attack and defense agents, and demonstrated cyber-physical evaluation capabilities (see Section 2.3 for a full description of Year 4 GATE plans). Beyond GATE, we continue collaborations with developers of other cyber ranges, such as USC SPHERE. For computation infrastructure, we continue to submit proposals to commercial providers such as AWS and Microsoft Azure, and pursue support from organizations such as the Center for AI Safety (CAIS) Compute Cluster [83], which provides a shared infrastructure featuring 240 NVIDIA A100 GPUs.

Risk 3: Rapidly changing landscape of AI and cybersecurity (continued risk). The landscapes of both AI and cybersecurity are rapidly changing. New AI models and capabilities are emerging at an unprecedented pace. Meanwhile, new cyber and cyber-physical attacks, vulnerabilities, and paradigms are also emerging. Without recognizing and addressing such new trends, ACTION's research may face the risk of losing its relevance and leading edge.

Mitigation: We have identified, and will continue to track, new research trends and opportunities on the horizon. The planned research activities presented in this S&IP reflect recent emerging trends in AI and security not fully addressed in our original proposal, including the application and robustness of LLMs, agentic programming paradigms, physical AI security, and the security of AI coding systems. We will continue to monitor future landscape changes and paradigm shifts and adapt our S&IP accordingly, based on inputs from the ACTION researchers, select EAB members, and the broader research and development community.

Risk 4: AI's negative impact on security and society (continued risk). The power of AI has been unleashed with unprecedented magnitude and speed, which has understandably caused increasing concerns about AI's potentially negative impacts on society. AI agents without sufficient human control may commit wrongful system alerts, suspend critical services, or perform misleading attributions. More seriously, the AI stack may be abused by adversaries to develop malicious agents that disrupt or counter the defense functions performed by our agents.

Mitigation: For each of the AI and SEC thrusts, we will explicitly determine the possibility of AI stack misuse and AI agent compromise, and develop the corresponding countermeasures. Each ACTION technology or solution will be accompanied by a misuse and compromise mitigation plan. We will also develop

theories and models in Thrust AI-4 with the aim of defining strategies and tactics that will, under certain conditions, provably give benign agents the upper hand during an engagement with malicious agents. Our ethical research policies for ACTION, including responsible disclosure protocols and ethics training requirements, are detailed in Section 13.

Risk 5: Integration lag in research and curriculum development (continued risk). As a large-scale, multi-institutional project, it is challenging to coordinate and synchronize integration efforts in both research and curriculum development, which may present the risk of integration lag.

Mitigation: The ACTION Executive Committee, which meets weekly, acts as the integration task force to coordinate core research and development tasks. For curriculum development, we have created an inventory of existing AI+Security courses and modules across all ACTION campuses. The NSF education supplement received in Year 3 has significantly accelerated curriculum integration by formally connecting UCSB, UW, Georgia Tech, and Western Washington University in a joint development effort. See Section 7 for a full description of the RFP and integration process.

Risk 6: Team turnover, role changes, and loss of key personnel (continued risk). With large projects and teams, there is always the risk that critical team members leave or take on additional responsibilities that make it difficult for them to fulfill their tasks for the ACTION Institute.

Mitigation: We have outlined a detailed Continuity Plan in Section 14.

Risk 7: Decentralized research among ACTION faculty.

Mitigation: The annual internal RFP process, launched in Year 2, has been the primary structural response to this risk. By requiring faculty to submit proposals that demonstrate alignment with the S&IP and cross-institutional collaboration, the RFP creates both accountability and investment. See Section 7 for a full description of the RFP process and its outcomes.

Risk 8: Narrow measurement of the success of the Institute.

Mitigation: We continue to work with our EAB to consider the bigger picture of ACTION. The RFP process allows us to measure which Institute-funded projects were most successful, most impactful, and most aligned with the Institute's vision. The

technology transfer table developed in Year 3 extends our success metrics beyond publications to include tool adoption, startup formation, and government engagement. The Google Scholar profile provides quantitative measures of scholarly influence, and the formal AI stack definition planned for Year 4 will provide a concrete architectural artifact against which external adoption can be measured.

Risk 9: Closing the SEC loop, completing the Security Lifecycle, and advancing all research thrusts equally.

Mitigation: SEC-4 has been substantially revised for Year 4, with clearer plans for closing the loop and clearer ties to all other research thrusts. Two Year 4 projects anchor the thrust: the Rollback Protocols project, which addresses semantic recovery of agent memory following attacks, and the Agentic Vulnerability Discovery and Remediation project, which delivers a full pipeline from vulnerability detection through patch generation and verification. Both projects include explicit GATE integration plans and connect to multiple upstream thrusts.

Risk 10: Lack of a formal AI stack definition.

Mitigation: Defining the AI stack in concrete terms is a primary deliverable for Year 4. The Year 4 mini-projects have been selected with this integration goal explicitly in mind, and the Year 4 narrowed scenario provides a concrete operational context against which the stack can be defined and validated. The GATE Collaboration Portal planned for Year 4 will serve as the shared repository where stack components are documented, versioned, and made accessible to all ACTION researchers and external partners.

Risk 11: AI and SEC thrusts evolving independently. This risk is distinct from the decentralization concern (Risk 7) and the SEC loop concern (Risk 9).

Mitigation: The Year 4 S&IP has been designed with this risk as an explicit organizing concern. The narrowed overarching scenario requires AI and SEC thrusts to work together to produce a complete defense. The AI+Cyber cohort structure, the RFP requirement for cross-thrust collaboration, and the GATE integration requirement for all Year 4 projects are structural mechanisms designed to promote convergence. The formal AI stack definition planned for Year 4 will create a shared technical vocabulary and architecture that makes the dependencies between AI and SEC thrusts explicit and measurable.

Risk 12: One-year research horizon sacrificing long-term vision.

Mitigation: This S&IP addresses this concern by adopting a two-year horizon for proposed research (Years 4-5), with each thrust describing research directions for both years. The SLED managed security services vision described in Section 6.5 and the Security as a Service model connecting ACTION to other NSF AI Institutes reflect a multi-year sustainability and impact horizon. The formal AI stack definition planned for Year 4 is itself a long-term artifact that will provide the architectural foundation for ACTION's contributions in Years 4, 5, and beyond. ACTION will continue to develop and articulate its long-term vision at each annual review, using the EC's strategic planning process to ensure that short-term project selection serves long-term goals.

6 Education and Workforce Development

In 2022, the United States boasted a workforce of over 1.7 million Software Developers, with a projected growth of 25% by 2032 [84]. This significant expansion stands in stark contrast to the average growth rate of all occupations, which hovers around 3%, underscoring the pivotal role of Computer Science in the job market. Additionally, the U.S. Bureau of Labor Statistics anticipates a 35% surge in positions related to Information Security [85], indicating a growing need for skilled professionals in cybersecurity and computer science fields. To meet these demands, the U.S. will require more students choosing to pursue careers in Computer Science. The obvious impact that AI will have on society only exacerbates the need for a job force proficient in technology.

To help meet these societal demands, the second of the ACTION Institute's three main goals is to educate students and the workforce about how AI and applications based on intelligent security agents can be used to protect the nation's critical infrastructure against attacks. More than this, we believe learning specifically about computer security can actually generate interest in Computer Science in general. In addition to developing and promoting cutting-edge courses that focus on cybersecurity and AI, and in addition to establishing immersive in-person research experiences at ACTION partner campuses, the Institute proposes novel education and training that revolves around Capture The Flag (CTF) challenges, which are puzzles aimed at teaching fundamental skills in cybersecurity [86].

Our strategy for improving education and workforce development has to vary according to the age and experience of our target population. High school students with limited experience with computer languages and coding would obviously require different materials than undergraduates considering a career in technology. And post-graduates employed in industry might or might not have experience with coding;

they might already be experts in the field, or they might be new hires. For this reason, ACTION proposes to focus education and workforce development on five distinct groups (Figure 3): (i) high school; (ii) undergraduate; (iii) early-graduate; (iv) advanced-graduate and post-doctoral; and (v) the emerging workforce.

6.1 K-12 & Community

ACTION has two primary modes for educating the pre-college community: participation in online security competitions and classes/training in computer science.

International Capture the Flag (iCTF). Our first educational tool is an annual security competition centered around AI and security and focused on K-12 students, high school students in particular. In the last decade, security competitions have become increasingly popular as a tool to broaden participation, attract talent, and motivate students to explore educational opportunities beyond traditional approaches (i.e., worldwide security competitions such as DEF CON's Capture The Flag [87]). UCSB has been organizing a yearly open security competition, called iCTF [88], since 2003, which has involved dozens of educational institutions and thousands of students around the world. The PIs have open-sourced the software to run the competition [89], which is also used in educational settings to organize educational security competitions.

In Years 1 through 3, we leveraged our faculty's prior experience to create a novel competition framework that focuses on usability and playability, accessible to high school educators and students who do not have sophisticated administration skills. We will continue offering these competitions during future years of ACTION, with each competition organized by a cross-institutional team of faculty members, undergraduate, and graduate students. The low barrier to participation will support the inclusion of students from a wide range of schools and communities, and we will devote special efforts to recruiting schools that serve students with limited prior access to cybersecurity education.

While CTF competitions may have a low barrier to participation, creating and managing the associated infrastructure requires skills and, for large competitions, the ability to scale. We will continue development of a turn-key solution that leverages the compute-on-demand capability of cloud-based infrastructures, shadowing our GATE research testbed, to both run our own competitions and, more importantly, allow others to use the framework to create their own competitions through an easy customization process. This approach will allow educational institutions that might lack the resources

to run a competition on their infrastructure to access an important and engaging educational tool.

In Year 3, we developed the Agentish platform, which teaches the basic tenets of agentic programming in a hands-on manner. Rather than solving challenges directly, students conceptualize a solution and then build agents to do the work for them, learning to develop and deploy agents and teams of agents to accomplish tasks with a focus on computer security. We launched Agentish as part of the iCTF 2026 "Battle of the Agents" competition and will continue to develop it in Year 4, making it scalable and broadly accessible, with the goal of deploying it as an important tool for education in agentic programming across a wide range of institutions.

Classroom training in computer security. The ACTION Institute's second mode for training the K-12 community involves synchronous and asynchronous training in CS, AI, and cybersecurity education. In 2024, we developed and launched a series of training modules based on actual iCTF challenges. In Year 4, we will continue to refine these modules and improve their presentation through YouTube videos and walkthroughs for the Agentish platform.

Our planned modules will walk participants through challenges similar to those being played in live security exercises, and will provide background information on challenges, including tips for solutions, historical and ethical implications of related vulnerabilities and exploits, and real-world cases where challenges would apply. Modules will be sequential and will grow in complexity.

To test the efficacy of our educational modules, we will continue our interaction with nearby high schools, helping to set up student clubs focused on Computer Security and AI and training programs for teachers. With iCTF traditionally held in late Fall, and sometimes Spring, we envision at least three months of opportunity to train students at the beginning of the school year. Our goal is approximately 12 courses/modules introduced on a weekly basis. In 2024, ACTION began this work with a local high school. In 2025, we began branching out to other high schools in the Santa Barbara and Chicago areas. We anticipate continued expansion of outreach in Year 4.

Potential for K-12 Cybersecurity Education Expanded Nationwide. ACTION recognizes that merely building infrastructure for training in CS and Computer Security is not enough. We also need to help build and engage the entire pre-college community of Computer Science enthusiasts. We have already begun this by connecting with various organizational resources and campus networks near our partner institutions,

using these connections to advertise iCTF at high schools near ACTION's twelve campuses, thereby creating a national community of participants.

ACTION has also begun to work with state and county-wide educational offices to further the Institute's training agenda. We started with Santa Barbara and will use our experience with local organizations as a template for expanding to partner campuses. The Santa Barbara County Education Office (SBCEO) has direct ties to each of the high schools in SB County, as well as to districts and high schools in neighboring counties through the California Mathematics, Science, and Computer Science Partnership [90]. Our hope is that a working relationship with SBCEO will ultimately lead to contacts throughout Southern California, and eventually beyond the state borders.

Santa Barbara County High Schools participating in ACTION activities have already noted increased interest in adopting cybersecurity training as part of college preparation. Leveraging our ties with DHS and CISA, we are working to expand plans for a national curriculum for cyber training to local high schools. ACTION is seeking opportunities to collaborate with CISA in this effort.

Engaging K-12 Teachers as Partners in Cybersecurity Education. A key Year 4 strategy for ACTION is to deepen its engagement with K-12 teachers, not only as recipients of training but as active partners in expanding cybersecurity education and vocational training at their schools. Norfolk State University demonstrated this model in Year 3 by including seven teachers alongside high school students in its first ACTION-sponsored CTF training camp. In Year 4, ACTION will build on this foundation by offering K-12 teachers more structured pathways for continuing education and professional development in AI and cybersecurity.

Specifically, ACTION will work with teachers interested in adopting cybersecurity education or vocational training programs at their schools, offering them opportunities to interface with relevant government and certification bodies. CISA's National Initiative for Cybersecurity Careers and Studies (NICCS) provides a dedicated portal for K-12 teachers that includes cybersecurity curricula, professional development workshops, and education tools. ACTION is well-positioned to connect teachers with these resources while supplementing them with ACTION-developed materials, including iCTF challenge modules, Agentish walkthroughs, and content from the Composable Curriculum. The College Board's Advanced Placement Cybersecurity course, which launched nationally in the 2026–2027 school year, represents another important touchpoint: ACTION will seek opportunities to ensure that ACTION-developed curriculum and tools are aligned with AP Cybersecurity standards,

so that teachers who participate in ACTION training programs are also prepared to teach college-credit cybersecurity courses.

K-12 schools are among the most frequently targeted sectors for cyberattacks, yet most districts are, in CISA's words, "target rich, cyber poor" [91]. Teachers who understand AI-driven cybersecurity concepts are therefore not only more effective educators but also frontline contributors to their institutions' security posture. By investing in teacher development, ACTION multiplies the reach of its educational mission: one teacher trained is potentially dozens of students per year, across multiple school years, in communities where ACTION does not have a direct physical presence.

Annual activity timeline addition:

- June: Teacher-training workshops offered near additional ACTION campuses, including connections to CISA NICCS resources and AP Cybersecurity alignment.
- August - September: Identify and recruit teachers interested in adopting cybersecurity vocational programs, with particular attention to schools near ACTION campuses in Santa Barbara, Chicago, Atlanta, and the Hampton Roads region.
- July - September: Refine training modules based on prior year feedback.
- August - September: Expand training program to more high schools through in-person outreach (Target = +10 high schools/year).
- September - October: Annual iCTF preparations. Involves challenge creation across Institute, platform development, website updates, and recruitment of high school teams.
- November - December: Host annual iCTF. Institute evaluator administers post-event evaluations and interviews.
- January - March: Promote team participation in other online security exercises, such as PicoCTF.
- January - May: Help develop new student clubs at more high schools participating in iCTF.

6.2 Undergraduates

ACTION's education and training plan for undergraduates has three components: online security exercises; undergraduate classes; and summer research experiences.

iCTF for undergraduates. Similar to our plans for high school education, we will offer an annual version of iCTF focused on college students. The undergraduate iCTF will be more challenging than the high school version and will assume more experience with Computer Science concepts. Advertising and promotion will be conducted through faculty connections across the Institute and beyond, as well as through campus computing and technology organizations at each ACTION institution. The Agentish platform will continue to be developed and utilized for agentic challenges.

Novel undergraduate curricula. ACTION Institute faculty will continue to collaborate to create novel undergraduate curricula centered around the concept of AI for cybersecurity that will be (i) taught in person at the Institute's organizations, (ii) provided as a blueprint with materials and supporting artifacts to other organizations to create similar curricula, and (iii) provided as Massive Open Online Courses (MOOCs). The use of MOOCs and packaged classes will support the education of students well beyond the limits of the Institute's organizations. We will leverage existing networks at the Institute's organizations, including connections to regional universities, community colleges, and institutions in the NSA's Centers for Academic Excellence in Cybersecurity community, for outreach and dissemination.

We will compile a list of undergraduate courses taught across all ACTION campuses in the areas of Computer Security and AI. We will assemble a committee of students through the Student Advisory Council and faculty to review these classes, both for overlapping and missing elements, and then brainstorm potential new classes accordingly. Our goal is to create classes that teach at the intersection of Computer Security and AI, rather than simply introducing new courses that remain siloed.

Undergraduate research internships. Participation in research is among the most significant of high-impact practices for student success, shown to be especially effective for the retention of students who might not otherwise have access to research opportunities [92, 93, 94], and is also seen by employers as valuable preparation for work [95, 96, 97]. All ACTION Co-PIs will be encouraged to onboard U.S. undergraduate students into their research labs.

The ACTION Institute will offer a summer research experience each year for eight to twelve undergraduate students to work on-site at an ACTION campus for a period of eight weeks, roughly late June until mid-August. Internships will involve immersion in a team-based investigation involving AI and computer security. General plans for internships will be conceived during the annual, internal RFP. ACTION faculty will use their mini-research proposals to announce their intent to include ACTION-sponsored

undergraduates in their research. In this way, ACTION leadership will be able to plan and coordinate the distribution of undergraduate interns across multiple Institute campuses.

When possible, we will employ a mentorship model where pairs of undergraduates are teamed with a Master's level graduate student as well as a senior PhD student and/or Postdoctoral researcher. Students will also receive professional development training in select management and entrepreneurial skills. At the end of each summer, interns will learn to prepare research posters and present their projects as part of a diverse, campus-wide undergraduate research symposium.

Undergraduate interns will be selected according to the impact the internship will have on the applicant's education, as well as the potential impact the applicant will have on any proposed research projects. The Student Advisory Council will play a major role in the initial rounds of student selection. All applications and associated materials will be managed through a password-protected portal maintained on secure servers at UCSB. Applications for summer internships will be advertised broadly, across the Institute, but also at other colleges and universities, with a particular focus on reaching students who might not otherwise have such an opportunity for research.

Summer interns will either receive housing through a university-owned apartment, or a stipend to help with living expenses. All undergraduate participants will participate in program evaluations and will be encouraged to think about how they can bring positive experiences back to their home institution.

Collaboration with Stakeholders. Not every undergraduate student will have access to a research internship, but ACTION's collaboration with NSF aims to ensure that opportunities for undergraduate skills development are as widely available as possible. Through NSF's connection with the National Security Agency (NSA) and the National Centers for Academic Excellence in Cyber Defense, ACTION has also identified an opportunity to assist nearly 500 campuses with their undergraduate programs, many of which are transitioning from a solely Cybersecurity focus to one at the intersection of Cyber and AI. ACTION is well-positioned to help with this transition.

Annual activity timeline, June - May:

- June - August: Summer research experience. Minimum of eight students per year.
- October - November: Faculty submit mini-proposals to ACTION RFP, announcing intent to include one or more undergraduates in next summer's research project.

- October - December: SAC collects lists of undergraduate courses across all ACTION campuses. Analyze available courses; prepare/plan additional courses.
- September - October: Annual iCTF preparations. Involves challenge creation across Institute, platform development, website updates, and recruitment of undergraduate university teams.
- November - December: Host annual iCTF. Institute evaluator administers post-event evaluations and interviews.
- December: Announce summer undergraduate research experience. Open online application portal.
- February - April: Select cohort of summer undergraduates. Involves participation from SAC as well as those faculty who will support undergraduate interns.
- April - May: Finalize project plans for summer research.
- Variable month: Launch new course and/or MOOC in AI+Cybersecurity, if developed.

6.3 Graduate Students

Our strategy for graduate student training and education mirrors the plans outlined above for undergraduates, but with a few differences. We will develop new graduate classes and implement them across the Institute. Also, participation in iCTF and summer internships remain; however, our goal is to provide opportunities for graduate students to assume additional responsibility through leadership. We also offer an institute-wide seminar series catered to the interests of graduate students.

Novel graduate curricula. Similar to the plans for undergraduates, the ACTION Institute will develop a series of advanced graduate classes focusing on the intersection of AI and cybersecurity that will complement any existing graduate curricula. In particular, the Institute will support the creation of courses that directly map into the eight research thrusts outlined in the ACTION research plan, highlighting the relationships between foundational AI and use-inspired research. Development of these courses will occur alongside the development of undergraduate courses. In Year 3, a novel graduate course, AVATAR (Autonomous Vulnerability Analysis, Triaging, And Repair), was developed and taught for the first time in Year 3, and will be offered again in Year 4 with agentic programming content added at the beginning.

iCTF for graduate students. We will not have a version of iCTF targeting Masters and PhD students. Instead, graduate students advised by ACTION faculty will participate in iCTF from the other side. They will develop challenges and help run the security

exercise. In the months preceding iCTF, graduate students in ACTION Institute Labs will collaborate on new challenges and then test them. The students will be empowered with running the competition, using the platform developed in prior years.

Graduate research internships. The ACTION Institute will run an annual graduate internship program in parallel with the undergraduate summer experience. This opportunity will allow four graduate students each year to spend a summer at one of the participating campuses, joining undergraduates and postdocs in an 8-week summer research experience. Applications will be advertised and managed in the same manner as the undergraduate applications. Selection criteria will consider background and the impact the internship will have on the applicant's education, as well as the potential impact the applicant will have on any proposed research projects. Our target graduate students will be those pursuing a Master's degree, or those who are just starting their PhD within their first or second year. We want to provide these students with a new research experience, something to catalyze their future efforts.

Where possible, the Institute will also facilitate summer internships with academic and industry partners for graduate students affiliated with ACTION.

Graduate seminar series. As part of graduate training, the Institute will provide a cross-institute seminar series, organized and led by graduate students under the direction of the Student Advisory Council. These seminars will focus on issues and topics of specific interest to graduate students and will include external speakers. The ACTION Director, Managing Director, and members of the Executive Committee will help attract and invite external speakers identified by the Student Advisory Council. We will aim for distinguished lectures on a quarterly basis, with other talks and presentations scheduled ad hoc.

Additional graduate training. All graduate students will be offered professional development training through existing programming at the Institute's organizations and will be required to take classes in the responsible conduct of research. By engaging in ACTION research, they will have access to the GATE infrastructure and the network of ACTION faculty and other personnel.

Annual activity timeline (June - May):

- June - August: Summer research experience. Minimum of four students per year.
- September - October: Annual iCTF preparations. Involves challenge creation across Institute, platform development, website updates, and recruitment of high school and undergraduate teams.

- October - December: SAC collects lists of graduate courses across all ACTION campuses. Analyze available courses; prepare/plan additional courses.
- October - May: Seminar Series. Distinguished lectures on a quarterly basis.
- November - December: Host annual iCTF. Institute evaluator administers post-event evaluations and interviews.
- December: Announce summer graduate research experience. Open online application portal.
- February - April: Select cohort of summer graduates. Involves participation from SAC as well as those faculty who will support graduate interns.
- April - May: Finalize project plans for summer research.
- Variable month: Launch new course and/or MOOC in AI+Cybersecurity, if developed.
- Variable month: Facilitate internships with industry.

6.4 Postdocs and Advanced/Senior PhD Students

In 2023 the ACTION Institute launched a postdoctoral fellowship program that would provide up to two scholars per year support for a six-month visit to one of the Institute's partnering organizations, where postdoctoral researchers would be able to directly collaborate with other members of the Institute. The fellowship was meant to create strong ties and foster collaboration among the Institute's sites while also exposing postdoctoral scholars to different academic environments and expanding their professional networks.

While our initial strategy was to leverage organizations such as the University of California President's Postdoctoral Fellowship Program for recruitment, ACTION faculty later resolved to limit the pool of applicants to those scholars already advised by partners in the Institute, with the goal of finding additional threads for collaboration.

Following an evaluation of the program, and at the suggestion of the Student Advisory Council, the Institute leadership has decided to expand the postdoctoral fellowship to include advanced graduate students who are nearing the end of their PhD. As noted by the SAC, postdoctoral scholars are often on a tight contract with their advisor and might not be able to take six months at a different location. Graduate students, on the other hand, are well-poised to extend their ongoing research through cross-institute collaborations, and such a collaboration might lead to a future postdoctoral appointment. Starting in 2025, we encourage advanced PhD students to also apply for the six-month exchange. All mention of training below applies to selected graduate

students in addition to selected postdocs. Our intent is to offer at least two of these collaborative opportunities each year.

Postdoctoral Fellowships. ACTION will manage applications for the postdoctoral fellowship program in the same manner as the applications for undergraduate and graduate interns. All data will be handled through secure servers. The postdoctoral fellowships will have a timeline with beginning and ending dates that are open-ended. Scholars will be asked to choose an advisor from the entire list of ACTION faculty, noting that the requested advisor must agree to host the applicant, as mentorship could require a significant amount of work.

Postdoctoral Researcher Mentoring Plan. The ACTION Institute PIs and faculty recognize the need to actively promote postdoctoral researchers' professional development and are committed to a program of structured mentoring activities. The goal of this program will be to provide the skills, knowledge, and experience to prepare postdoctoral researchers to function independently as scientific professionals. Following the guidance of the National Academies of Sciences, Engineering, and Medicine [98, 99], we will offer our postdocs a range of opportunities to learn or improve a number of critical skills.

- **Orientation:** On-boarding will start with an orientation consisting of in-depth conversations between the PI and the postdoctoral researcher. Mutual expectations will be discussed and agreed upon in advance. Orientation topics will include (a) the amount of independence the postdoctoral researcher requires, (b) the university culture of collaboration and interaction with coworkers, (c) productivity, including the importance of scientific publications, (d) work habits and laboratory safety, (e) documentation of research and data management to ensure that the work can be shared with and continued by other researchers, and (f) expectations for the responsible conduct of research.

- **Career Counseling:** Counseling will be directed at providing the postdoctoral researcher with the skills, knowledge, and experience needed to excel in their chosen career path. In addition to guidance from their PI, postdoctoral researchers will be encouraged to explore and discuss career goals with faculty, industry, and other professionals affiliated with the Institute. The ACTION Institute organizations are members of the National Postdoctoral Association (NPA), and our postdoctoral scholars will have access to NPA resources and the national community.

- **Experience with Preparation of Grant Proposals:** Postdoctoral researchers will have an opportunity to learn best practices in proposal preparation including identification of key research questions, definition of objectives, description of approach and rationale, and construction of a work plan, timeline, and budget.

- **Publications and Presentations:** The postdoctoral researcher advisor will provide guidance and training in the preparation of manuscripts for scientific journals and presentations at conferences. In addition, postdoctoral researchers will have access to workshops on related topics regularly offered at our campuses.

- **Teaching and Mentoring Skills:** The postdoctoral researchers will have the opportunity to develop teaching and mentoring skills by participating in the research meetings within the PI's research group, as well as the ACTION Institute's regular sync-up meetings. Postdoctoral researchers will also be offered the possibility to develop their teaching and mentoring skills by participating in some of the numerous educational and outreach activities organized by the Institute.

- **Success of the Mentoring Plan:** The effectiveness and progress of our mentoring plan will be assessed by monitoring the personal progress of each postdoctoral researcher toward their stated goals during their application and orientation. The evaluation team will conduct regular interviews with all postdocs to assess satisfaction with the mentoring program and track progress toward career goals after finishing the postdoctoral research experience. Evaluative findings will be included in quarterly and annual reports. The ACTION Director will also meet with the postdoctoral fellows regularly to help assess the program.

Graduate Student Researcher Mentoring Plan. Considering the overlap of postdoctoral scholars with advanced PhD graduate students, ACTION PIs have adopted an optional rubric for use in measuring the growth of their graduate students. Developed for an NSF IGERT program (award# DGE-1258507, PI Ambuj Singh) by a Professor of Sociology, the rubric helps track student advancement and potential for impact across three categories:

- **Professional Development:** Graduate students will (i) develop effective oral presentation skills; (ii) develop familiarity with and knowledge of post-graduate career options; (iii) learn how to, and actively participate in, publishing professional articles; (iv) develop leadership and mentoring skills; (v) gain a

grasp of ethics in research and scholarship; (vi) develop skills in communicating science to the public.

- **Learning to use the Scientific Method:** Graduate students will (i) locate research within the scientific literature; (ii) identify and formulate useful scientific questions; (iii) specify and test useful hypotheses; (iv) identify useful new research opportunities and/or applications of one's research.

- **Working in an Interdisciplinary Environment:** Graduate students will (i) become effective at hearing and understanding scientific discussions that are relevant and up-to-date in disciplines other than their own; (ii) develop a familiarity with using research tools from disciplines other than their own; (iii) gain global and practical perspectives; (iv) build innovation skills.

ACTION does not require use of the rubric. Rather, we offer it as an optional tool for mentoring.

6.5 Workforce Development

To support workforce development, ACTION will look to augment the tools we design for educational institutions to also be used for companies and non-profit organizations, to create novel competitions that support the education of the workforce around the themes of AI and cybersecurity.

The PIs have substantial experience in organizing and participating in these kinds of competitions [88, 89, 100, 101]. In future years, the ACTION Institute will build on this expertise to develop tools that will help organizations and companies create a systematic talent pipeline at all professional levels, using competitive learning and training to attract, engage, and empower candidates who will become the next-generation AI/cybersecurity workforce. By publishing novel educational challenges that can be used in a workforce development setting, our influence will extend outside the institute and scale to a broader set of stakeholders who lack tools for training the workforce in security and AI.

ACTION aims to leverage existing industry and government partnership programs, such as CISA's Skilling Academy, to help design workforce development tools, learning modules, and short courses, with customization to fit specific industry sectors, job types, and skill levels. We have also taken note of NSA's community of schools through

their Center of Academic Excellence in Cyber Security (CAE) initiative. Of the nearly 500 CAE schools, slightly more than 100 programs, with support from NSF, are transferring to an AI+Cybersecurity curriculum. ACTION is well-poised to help with both the greater CAE initiative and the AI-influenced subset of schools by helping the managers of these programs prepare their graduates for careers in cybersecurity.

We will also continue to reach out to other community colleges and schools with cybersecurity vocational programs to see if our tools for workforce development are in line with their curricula.

Understanding the Long-Tail of the Cybersecurity Workforce. A final area under consideration for ACTION's impact on workforce development involves an understanding of these three noticeable trends:

1. **Agencies like CISA and DHS are increasingly shifting cybersecurity responsibility to State, Local, and Education (SLED) organizations.** But these entities are chronically under-resourced and unprepared to defend themselves.
2. **Nation-state actors are no longer just targeting prime contractors.** In addition, they are targeting the forgotten supply chain: agricultural cooperatives, rural manufacturers, and regional infrastructure. This represents a persistent cyber engagement with asymmetric consequences.
3. **Tomorrow's workforce must be AI-native.** Students and professionals who cannot leverage AI for analysis, defense, and decision-making will be left behind in an economy defined by automation and augmentation.

When considering these three trends as a whole, it becomes evident how AI is about to transform the economics of cybersecurity. Just as cloud computing redefined cyber infrastructure, AI will soon redefine the costs behind cyber defense, along with which businesses and organizations can actually afford it.

ACTION is well positioned to impact this national concern. ACTION can become a training ground for AI-native Cyber Defenders.

Although it will take considerable planning, there remains a future potential for ACTION to launch applied AI as a managed security service provider (MSSP). If it could develop real-world, AI-augmented MSSP services delivered by students and faculty, helping SLED organizations secure themselves while giving students hands-on experience, ACTION could shift from a grant-dependent model to a sustainable service model that offers training while also defending.

Considering the economic impact of the long tail, we see a Security as a Service (SaaS) model as a potential collaboration with other National AI Institutes. Many of these institutes are already creating an AI-first academic and professional cyber workforce in niche domains like agriculture, manufacturing, water, and energy. ACTION could connect with all other AI institutes and beyond to become a national center of excellence for securing this long-tail industrial base.

This vision for a potential future for ACTION requires additional planning and thought. We include it in this S&IP because it emerged from a meeting with members of our External Advisory Board. We will continue to refine this and other strategies for institutionalization and national impact over the next two years.

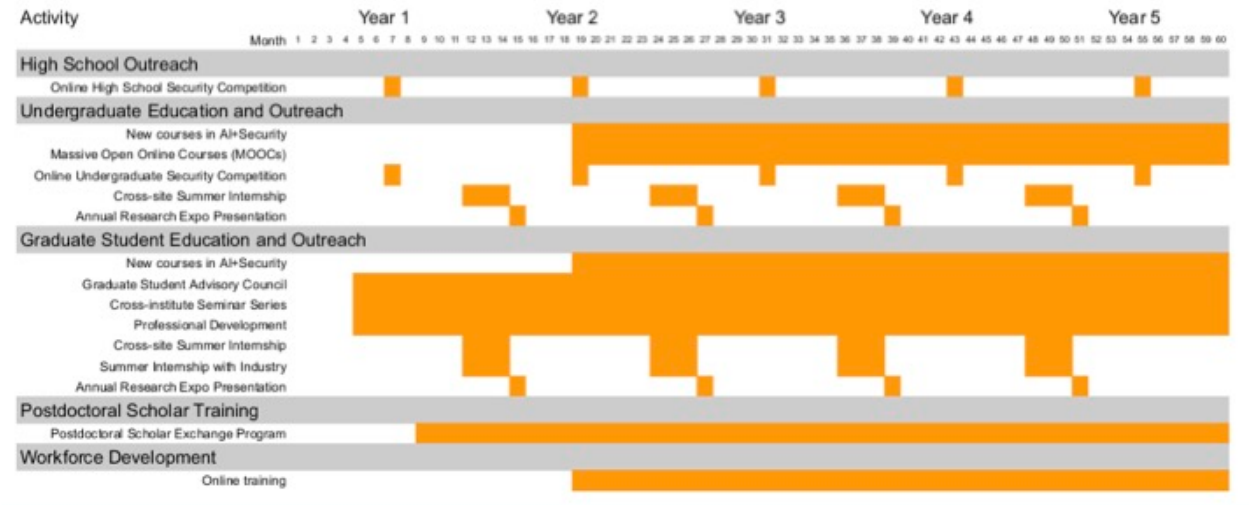


Figure 3: Timeline highlighting some of the milestones associated with the education and workforce development efforts.

6.6 Additional Education-related Projects, Year 4

ACTION will continue a number of education-based projects throughout Year 4 (June 2026 - May 2027) and beyond. These projects have the potential to impact all five targeted groups. As with the research-oriented projects, we will provide quarterly updates on these efforts.

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

Abstract: Cybersecurity is undergoing a rapid transformation due to the integration of machine learning (ML) and generative AI (GenAI) into components of the critical infrastructure. Threats now arise not only from traditional attack vectors, such as remote code execution and authentication bypasses, but also from model-specific vulnerabilities, such as adversarial examples, data poisoning, model leakage, and prompt injection. Existing curricula, often static and siloed, fail to reflect these evolving realities. There is a critical need for a flexible, modular, and hands-on curriculum that covers these advanced concepts at the intersection of AI and Cybersecurity and promotes deep understanding through self-guided, inquiry-based learning supported by AI tutors.

Proposed Curriculum: Course modules are designed to be independently useful but composable, enabling offerings at the undergraduate, graduate, or professional level depending on need and audience. Each week-long module consists of lectures accompanied by a hands-on lab, and each topic can be offered as a full 10-week course or as shorter mini-courses. All courses stem directly from research carried out by the ACTION Institute team. The six proposed course modules are:

1. **Machine Learning for Cybersecurity (ML4Cyber):** Weekly modules each focus on a specific security problem, including email spam detection, CAPTCHA solving, credit card fraud detection, and adversarial machine learning. Students implement defense strategies using real-world datasets.

2. **AI Alignment for Safe Downstream Applications:** An introduction to aligning LLMs with human values and strengthening model resilience. Topics include high-quality alignment dataset generation, jailbreak attacks, and inference-time defenses that mitigate attacks without introducing computational overhead.

3. **AI for Edge Intelligence (TinyML):** A hands-on course on deploying high-performing ML models on resource-constrained edge hardware. Modules address applications including wake word detection, sign language interpretation, and activity detection.

4. **AI/ML-based Security Analytics:** An introductory cybersecurity class examining the security lifecycle and appropriate AI/ML-based solutions. Students learn to formulate security problems as classification tasks, select and evaluate ML algorithms, and curate training data.

5. **Robustness of AI to Adversarial Manipulation:** An advanced course covering evasion attacks on detection models and data poisoning attacks on training pipelines. Students build skills in evaluating and hardening AI/ML-based security systems against adversarial manipulation.

6. **Iterative Approaches to AI/ML Security:** A practicum where students build a systematic toolbox for evaluating and improving AI/ML-based security solutions, iterating across attack and defense strategies using state-of-the-art adversarial frameworks.

Pedagogy and Assessment: The curriculum emphasizes inquiry-driven and experiential learning. Technologies adopted to facilitate such learning include Jupyter notebooks for hands-on labs, asynchronously delivered video lectures, and GitHub distribution for open access. Each course includes formative assessments through continuous evaluation during lab sessions, and summative assessments through course projects and student presentations.

AI Tutor: The curriculum will be supported by an AI-powered tutor that functions as a dynamic, interactive learning companion. Rather than dispensing direct answers, the tutor uses the Socratic method to prompt students with guiding questions, encouraging them to explore and discover their own knowledge. The tutor can work through examples of varying difficulty levels and generate variations of sample quizzes. This builds on the Lowkey Genius chatbot developed by Dr. Poovendran's team in Year 3 as an early deliverable of the NSF education supplement.

The curriculum is being disseminated through multiple channels. ACTION proposed incorporating Agentish CTF challenges into the NSF Faculty Development Summer Institute on AI in Cybersecurity Education, hosted by George Washington University and the University of Hawaii at Manoa in Summer 2026, to extend the reach of ACTION's agentic programming curriculum to a national cohort of cybersecurity educators. Through NSF's connection with NSA's Centers for Academic Excellence in Cybersecurity, the curriculum will also be made available to the more than 100 CAE programs transitioning to an AI+Cybersecurity focus.

Thrusts: Education and Workforce Development, AI-2 (Thread 3)

Project: AI and Security MOOC

Abstract: Building on our previous MOOCs, LLM Agents and Advanced LLM Agents, and their accompanying global hackathons, we propose a new, timely course at the intersection of two transformative domains: AI and Security. With the rapid growth of AI technologies and the increasing complexity of cybersecurity challenges, there is a critical need to both secure AI systems and harness AI to strengthen security. This new MOOC will explore these dual themes: AI for Security and Security for AI.
Thrusts: Education and Workforce Development

6.7 Assessment and Evaluation

ACTION will continue to work with its external evaluation team (UCSB Center for Evaluation & Assessment) to assess and evaluate the educational activities of the Institute. The evaluators prepare and administer surveys to ACTION's educational participants. They reach out to everyone who competes in the high school and undergraduate iCTF and meet with summer research interns. They compose informational reports using the collected data which they provide to Institute leadership for purposes of assessment. They also maintain a logic model representing the various goals and potential impacts of the Institute, and provide guidance and insight on student applications and the design of activities. The evaluation lead meets monthly with the ACTION leadership to check progress and regularly attends faculty meetings.

6.8 Education Risks and Mitigation

Risk 1: Lack of participation in iCTF. The concept of a CTF competition might be new to high schools. Students with limited experience in computer security might be reluctant to sign up to participate. A lack of participants could skew evaluative data and has a negative impact on ACTION's overall outreach and impact.

Mitigation: Outreach remains crucial if participation is meant to grow. ACTION must do the work in contacting schools about the iCTF opportunity. We use campus networks and faculty connections at each ACTION institution to help with this outreach. For more, see the section on Broadening Participation. Initially, ACTION graduate students and faculty will spend time, in-person, at high schools to help set up student-led hacking clubs that can prepare students for participation in security exercises like CTF. ACTION aims to prepare a curriculum that can be replicated at other high schools. In Year 2, ACTION sought cash prizes from industry sponsors which could be given to

winning high schools. We aim to continue this practice. Although iCTF is one of ACTION's educational deliverables, we recognize there are other CTF products. PicoCTF, hosted by Carnegie Mellon University, is the leading CTF for high school students. We encourage teams participating in iCTF to continue their training with other CTFs.

Risk 2: Lack of application pool for internships and fellowships. Our goal should be to have an application pool that is at least three times the size of our number of award recipients. The pool should also be broadly representative of students who stand to benefit from the experience.

Mitigation: ACTION has already engaged in, and will continue to engage in, outreach and advertising through personalized emails to targeted schools. We keep the website as up-to-date as possible. We hope to prepare a Strategic Communication Plan in the coming year, where various press releases and other Institute informational material can be kept for distribution.

Risk 3: Narrow reach of education and outreach. See Section 8, Risk 5 (Concentration of broadening participation activities on a few institutions) for a full discussion of this risk and its mitigation.

Risk 4: Who is ultimately responsible for high school education? See Section 8, Risk 4 for a full discussion of this risk and its mitigation.

Risk 5: Is CTF Dead? See Section 8, Risk 7 for a full discussion of this risk and its mitigation.

7 Integration

ACTION's goals for research, education, and knowledge transfer integrate into a single, coherent program. Our activities for broadening participation and outreach draw directly from the research. Through internships, seminars, and training activities, ACTION connects researchers throughout high school, college, and industry. Beyond the dissemination of research through academic papers, our GATE infrastructure provides novel opportunities for knowledge transfer.

7.1 ACTION's Integrated Research Approach

We foster integration across the axes of AI and cybersecurity through three collaborative vehicles: topics, teams, and applications.

Systematizations of Knowledge: Topic-centered integration, represented in Figure 4 as "early integration," sets the pace for our research and provides a common understanding of specific fields. We pursue "Systematization of Knowledge" (SoK) research threads that focus on the application of AI to cybersecurity tasks. Specifically in Year 4, we aim to further develop SoK threads on (i) AI for Vulnerability Analysis, (ii)

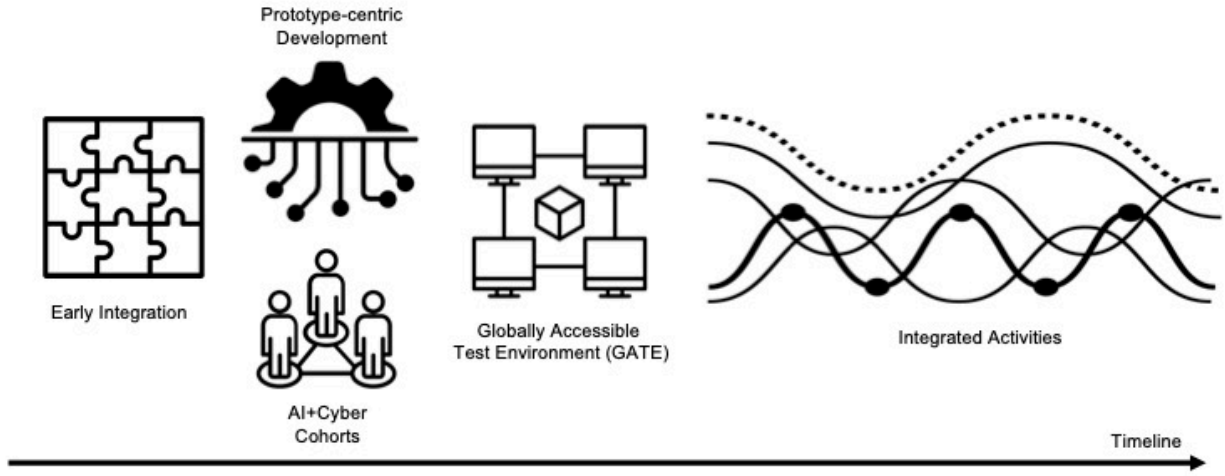


Figure 4: The ACTION Institute vision for integration

AI for Intrusion Detection, (iii) AI for Threat Intelligence and Attribution, and (iv) AI for Automated Response, the four points along our cybersecurity axis as described in Section 2. These SoK-focused research efforts will initially work on gathering existing research and systematizing the field, creating ontologies, identifying sub-fields of research, and highlighting research opportunities that have not yet been explored. These SoK threads will be led by graduate students with mentoring from the Institute's faculty. This will create an opportunity for students to better see the "bigger picture" of the application of AI techniques to cybersecurity while identifying interesting new lines of research. The SoK threads will also allow us to unite multiple institutions toward common goals, making inter-institutional and interdisciplinary collaborations intrinsic to the Institute.

The thread for AI for Threat Intelligence and Attribution began in Year 2 and was further developed and refined into an academic paper in Year 3. Faculty and students from U Virginia, UC Santa Barbara, UI Urbana-Champaign, Georgia Tech, and U

Chicago have all been active participants in the project. This SoK helped ACTION rethink and re-strategize Thrust SEC-3, moving from mere attack attribution into the broader idea of threat intelligence. In a similar manner, we anticipate this and future SoKs to also help with our course development.

SoK also provides ACTION graduate students with opportunities to explore new directions for research. Specifically, in Year 3, we began to look at implications of AI for cybersecurity in space, with satellites as the critical infrastructure requiring defense. Physical AI is another topic of interest. We anticipate the Student Advisory Council to further look into these two topics over the next year of operation.

Research Teams and Focused Projects: To integrate the research efforts of the Institute's members, we will continue the practice, launched in Year 2, of defining tailored mini-projects that span multiple research thrusts and threads and drive the research activities described in Sections 3 and 4 of this S&IP. These collaborative and focused projects will allow the Institute to devote resources toward specific prototypes or areas of major interest. Importantly, they will be carried out by AI+Cyber Cohorts, which are teams of researchers from both the AI and cybersecurity fields, who will work together to ensure that foundational AI research and use-inspired security applications are intertwined. The 19 projects to be undertaken in Year 4 are described in detail in Sections 3 and 4, each clearly articulating how they integrate and connect the various research thrusts of the institute.

Application and GATE: Our GATE framework, which is available to all ACTION researchers, will provide additional guidance and support for the integration of various research activities toward a common application. In Year 4, the overarching scenario involving coordinated attacks against chemical and water-processing plants with homogeneous IT infrastructure will serve as the primary integrating application context. In later years, we foresee the Institute's AI+Cyber cohort projects introducing new prototype applications that demonstrate how the AI stack is used by intelligent agents in carrying out tasks that are part of the cybersecurity lifecycle. To facilitate these projects, we maintain a code repository for the AI stack, to be cooperatively developed along the four AI thrusts, and strongly encourage efforts in the four SEC thrusts to adopt and apply the AI stack, either as a whole or selectively by layer.

GATE will serve as the designated evaluation platform for all institute-developed tools and agents. We will establish a clear and standardized integration path for research tools, ensuring that thrusts can effectively use GATE to validate their work in complex, multi-thrust scenarios. As one example, consider a Year 4 AI+Cyber cohort that includes

an expert in intrusion detection, an expert in threat intelligence, an expert in agentic systems, and an expert in large language models. This cohort proposes a project that delivers a novel approach to leveraging LLMs to create a human-agent interactive system, in which agents that identify and correlate alerts across the chemical and water-processing plant networks present to a SOC analyst an analysis of a detected multi-stage attack, and then use reasoning and knowledge about the domain to come to agreement on an intrusion abstract pattern that can be shared across facilities. This application would use capabilities provided by Thrust AI-1 (knowledge and reasoning), AI-2 (human-agent interaction), AI-3 (multi-agent collaboration), SEC-2 (detection), and SEC-3 (threat intelligence sharing), and would be validated end-to-end in GATE against the Year 4 overarching scenario.

GATE also supports ACTION's education and knowledge transfer missions through the Collaboration Portal, a lightweight "GATE-in-a-box," and an integrated CTF platform (Section 2.3).

Selection of Integrated Activities: Our process for identifying and implementing integrated activities through focused mini-proposals began in Year 2. Following ACTION's annual review in June, after a complete assessment of Institute activities and future direction from NSF, the ACTION EAB and other stakeholders, the Executive Committee and research thrust leads refine the S&IP and identify short-term priorities for the next year of activity. The EC then announces a request for proposals (RFP) from ACTION AI+Cyber Cohorts. The internal RFP defines a research budget and the number of expected awards. Research teams consisting of at least two faculty from different ACTION campuses then submit to this RFP. Each proposal is required to identify contributions to the relevant research thrusts as well as impacts to the Institute. Proposals are also required to include a budget justification and a description of how the research would be integrated into GATE.

Accounting for conflicts of interest, the EC and thrust leads then review all proposals. Those that do not meet the requirements for integration and collaboration are rejected. The integrated activities that represent each of the Institute's research thrusts, as well as further the Institute's goals for education and workforce development, are selected. By creating focused, short-term projects with clear achievable goals, the Institute is able to select the activities that best follow through on our integrated approach. At the same time, ACTION leadership is actively able to re-assess and re-focus the trajectory of the Institute. Our strategy for maintaining integration is to repeat this process in future years.

Annual activity timeline, June - May:

- June: AI Institute Site Visit and Annual Review.
- July: Review/assess prior year's activities. Feedback from NSF and stakeholders.
- August: Respond to feedback. Refine S&IP. Prepare themes for next year's RFP.
- November: Mini-proposals accepted.
- January: Integrated projects selected.
- February: S&IP revised to include new projects. Submitted to NSF with annual report.
- June: New projects begin.

Other Means of Integration and Collaboration: In addition to working toward integration of research and education, the ACTION Institute also promotes a collaborative and interdisciplinary culture among all participants. This culture creates connections and cross-pollination that will last beyond the Institute's lifetime.

The ACTION Institute uses Slack and email for day-to-day operations and organizes regular monthly virtual meetings for each research thrust as well as quarterly virtual meetings for all collaborators. The ACTION Institute also organizes a yearly in-person meeting to showcase results and provide opportunities for networking and exploring future research directions. The yearly meeting also gives an opportunity to the EAB to provide feedback through hybrid participation.

In Year 3, the Student Advisory Council introduced a new practice to promote integrative activities: an opportunity for SAC members to travel to another ACTION campus of their choice to present their research. This program was devised by the SAC in February 2025 and launched in Year 3 as a means to increase awareness among the ACTION community. The students are highly supportive and see the program as an opportunity for networking and professional development, in addition to fostering greater communication within the Institute.

Other collaborative and integrative practices, including cross-site internships, the Postdoc/Senior PhD Scholar Exchange, graduate student seminars, and SAC activities, are discussed in detail in the section on Education and Workforce Development.

7.2 Risks to Integration and Mitigation

Risk 1: AI and SEC thrusts evolving independently without converging on an integrated AI stack. See Section 5, Risk 11 for a full discussion of this risk. The RFP

process, the GATE integration requirement, and the formal AI stack definition planned for Year 4 are the primary structural mitigations.

8 Broadening Participation

All ACTION Institute personnel believe in the importance of ensuring the voices of technology users are also represented by the voices of the technology creators. For this reason, the ACTION Institute is dedicated to broadening participation in the areas of AI and computer security, as established in the Institute's main goal for education and training in this plan's Intellectual Theme (Section 1).

ACTION's approach to broadening participation is grounded in NSF's core principle that the benefits of scientific discovery should create opportunities for all Americans everywhere. Our strategy emphasizes expanding access based on geography, institutional type, socioeconomic background, and career stage. These are characteristics that have historically determined who does and does not have access to cutting-edge research in AI and cybersecurity. All of ACTION's broadening participation activities are open and available to any interested participant, with outreach designed to reach communities and institutions where such opportunities have been scarce.

The geographic distribution of ACTION's twelve campuses, spanning the Pacific Northwest, the Mid-Atlantic, the Southeast, the Midwest, and both coasts, facilitates the implementation of this strategy. The ACTION Institute's role as a nexus point for collaborative efforts will be framed by a "network of networks" for engagement in and dissemination of our research and educational activities.

The foundation of ACTION's strategy for broadening participation lies in the proliferation of AI and in the tangibility of computer security. The general population is increasingly aware of AI, simply based on the attention given to advances like large language models. Everyone arguably now has access to the technology. Computer security is also a hot topic, often associated with ransomware attacks, phishing scams, and identity theft. Everyone with a smartphone, or even with basic access to the Internet, has some level of concern and knowledge about cyber-threats and computer security.

Early exposure to Computer Science is crucial for broadening participation in the field [102, 103, 104, 105]. For this reason, we shaped our Education and Workforce Development activities to provide us with access to K-12 education. We envision our focus on cybersecurity exercises helping increase participation among young students at

high schools near ACTION campuses, particularly those in communities where such opportunities have historically been scarce. High school teachers also play an important role in mentoring and motivating young students [106, 107, 108]. We held our first instructor clinic in June 2025 and plan to expand this program to additional ACTION campuses in Years 4 and 5. Outreach through iCTF will help us move beyond the borders of the Institute.

Our broadening participation strategy includes the use of community organizations for help identifying schools where our programs can have the most impact. In California, we have leveraged the Early Academic Outreach Program (EAOP), a state-wide college preparatory program sponsored by the University of California that serves more than 30 high schools in 13 counties throughout Northern and Central California, prioritizing schools in communities with limited access to college preparatory resources. In Georgia, we will partner with Georgia Tech's Constellations Center for Equity in Computing, which works with Atlanta Public Schools to bring computer science education to a district where nearly 75% of students qualify for free or reduced lunch, a concrete measure of socioeconomic need. Similar community partnerships exist near most of the Institute's organizations.

The above plans for broadening participation highlight our intended efforts at the high school level. We also plan to broaden participation in AI+security (and Computer Science) among college students. Undergraduate and graduate outreach will leverage existing networks at the Institute's organizations, including connections to regional universities and community colleges that serve students from a wide range of socioeconomic and geographic backgrounds. With the help of the Institute's Student Advisory Council, we will also engage campus organizations focused on computing and technology at each ACTION institution, as well as existing summer undergraduate research programs that aim to reach students who do not readily have access to research opportunities. Through NSF's connection with NSA's Centers for Academic Excellence in Cybersecurity, a community of nearly 500 campuses nationwide spanning community colleges, regional universities, and institutions in every state, ACTION has identified one of the most significant geographic and institutional multipliers available to the institute.

ACTION's other education plans (MOOCs, streaming lectures) and summer research internships are also part of our goals for broadening participation. The online classes will provide novel instruction to students who might not otherwise have access. The internships will specifically target students who do not readily have opportunities for participation in research (see Sections 6 and 11 for more details). Recruitment for all

programs is conducted through open, competitive application processes reviewed by ACTION management and the Student Advisory Council, with selection based on fit and readiness to benefit from the experience.

Complementing early exposure and mentorship, research shows community development to also be an important requirement for recruiting, retaining, and broadening participation [109, 110, 111, 112, 113]. The establishment and long-term maintenance of an interactive, stimulated community of computer science enthusiasts, all connected to, and benefiting from, the research underway in the ACTION Institute, is probably the most vital tactic in broadening participation strategy. All new ACTION participants are welcomed into this community and introduced to students and faculty through Slack and Discord communications and attendance at Institute meetings and engagements. The end of an internship or advisory council appointment does not, by default, end participation in this community.

Annual activity timeline (June-May):

- August - October: Refine educational CTF modules for outreach.
- August - October: Implement in-person outreach: at least ten more high schools each year near ACTION campuses.
- November - December: iCTF, with participation from high school and undergraduate teams nationwide.
- June - August: Host summer internship, new cohort.

8.1 Broadening Participation Risks & Mitigation

Risk 1: Lack of evidence-based principles. There is a risk that the educational and outreach activities of the Institute were not designed with evidence-based principles around broadening participation in mind, leading to risks that these activities will not engage and advance students from communities with limited prior access.

Mitigation: ACTION leadership conducted a literature review of evidence-based research to evaluate and then either adjust or reaffirm the activities. Based on 14 articles, each with varying evaluation methods, both qualitative and quantitative, we found the greatest impact on broadening participation is to build a supportive community, and to provide active, responsive mentorship. Accordingly, we make regular use of platforms like Discord and Slack and make every attempt to increase awareness of our participants.

A second factor of success in the literature was early exposure to a topic (e.g., Computer Science or programming). That considered, we have increased our attention toward high school participation, particularly at schools in communities with limited prior access to research-grade technology education.

However, we have found evidence that incorporating LLMs like ChatGPT into a structured, hands-on learning environment, as we are doing by adding AI teaming to iCTF, can increase the efficacy of learning the material [114]. Maintaining academic integrity can become problematic when students do not understand the value of learning, but careful construction of the iCTF challenges can mitigate these issues. The students are, after all, learning to problem-solve; coding is a secondary skill.

Risk 2: On-going vs. one-time experiences.

Mitigation: We have intentionally worked at strengthening the graduate student and undergraduate communities within ACTION. We seek to increase the longevity of participation within these communities. For example, while summer internships might officially end in August, the collaboration can continue. This is especially the case when participants seek to draft a manuscript for conference submission, or to extend their work into new areas. Another example of prolonged participation in community exists within our communication platforms, where students who have graduated are still regular contributors to content.

Risk 3: Sustainability of outreach efforts. With high school curricula already set in many schools, there is a risk that outreach activities cannot scale without dedicated staffing.

Mitigation: ACTION will dedicate funds to appoint an Education Coordinator. At the end of Year 3, we have begun to better define the duties of this role, drawing on similar positions at other National AI Institutes. We see an opportunity for novel research in the field of Education, especially with the inclusion of cybersecurity curricula at the K-12 level, and are currently exploring the balance between administrative and academic activities in this role with personnel in UCSB's Department of Education.

Risk 4: Who is ultimately responsible for high school education? With high school curricula already being set in many schools, and with the pacing of many courses being designed to prepare students for testing at the end of the year, there is a risk that adding new material to a course might be implausible. The mitigation in many cases is for extra material to be introduced in a student-led, extracurricular club. Most of these clubs

require the support of at least one faculty member; but with high school faculty being already overwhelmed, the responsibility for leading clubs often falls on students. But students are often also overwhelmed and overscheduled with homework, after-school activities, college applications, and family commitments.

Mitigation: To run a successful high school program, there needs to be a well-organized charter in place for the club, with dedicated leadership. There also needs to be an accessible curriculum and an incentive for wanting to work through this curriculum. We are still fine-tuning these elements of working with high school students. Our successes thus far have resulted from building a community on campus and connecting it with the community at the university that is already thriving.

Risk 5: Concentration of broadening participation activities on a few institutions. Broadening participation efforts remain somewhat concentrated on a few institutions, primarily UCSB and NSU, and that ACTION should work to increase the breadth of outreach by involving a wider range of institutions and student organizations.

Mitigation: The NSF education supplement received in Year 3 represents the most significant structural response to this concern, formally extending ACTION's educational activities to the University of Washington, Georgia Tech, and Western Washington University. These additions bring new geographic reach and new student populations into the institute's orbit. In Year 4, Georgia Tech and UW will join UCSB and Rutgers as summer research host campuses, doubling the number of hosting institutions. UIUC has begun K-12 outreach in the Chicago area, and Rutgers continues to grow its summer research activities through the DIMACS REU partnership. ACTION's strategy going forward is to leverage the networks of its stakeholders, including IBM, DHS/CISA, and NSA's CAE community, to reach institutions and communities that ACTION cannot serve directly. Fewer, stronger points of contact with these networks will allow for more effective and sustainable outreach than attempting to distribute responsibilities equally across all twelve campuses.

Risk 6: Insufficient tracking of broadening participation impact beyond the immediate ACTION community. More broadly, ACTION's current evaluation infrastructure is most fully developed for activities where participants are known to the institute, such as summer interns and iCTF competitors. The reach of tools like zero2hero, the Agentish platform, and future MOOCs extends well beyond this known population, and without systematic tracking, ACTION cannot demonstrate the full scope of its broadening participation impact.

Mitigation: The AI+Cybersecurity MOOC has been designated as a Year 4 deliverable, which creates an opportunity to address this risk proactively rather than retroactively. As course development begins in Year 4, ACTION plans to involve the external evaluation team from the ground up, with the goal of designing outreach and tracking mechanisms into the course from the outset rather than adding them after launch. For tools already deployed, such as the Agentish platform and zero2hero, we will work with the evaluation team to develop lightweight instruments that can capture engagement data from users outside the formal ACTION community. Our intent is to measure reach across institutional types, career stages, and geographic regions, consistent with NSF's current priorities for creating opportunities for all Americans everywhere.

Risk 7: Is CTF Dead? The use of AI has changed the game of Capture the Flag. Frontier AI models can solve traditional challenges in a fraction of the time humans used to require. AI proliferation has the organizers of iCTF wondering about the future of such exercises, and whether people need to learn the low-level skills involved in developing solutions.

Mitigation: Anticipating the use of agents to solve CTF problems, ACTION spent much of Year 3 devising a platform that focuses more on agent development and deployment. If agents are the future problem solvers, then the design and utility of the agent has to be our focus. The platform, known as Agentish, was built into iCTF 2026, where participants, instead of solving a challenge on their own, had to design an agent, and in some cases, teams of agents, to solve the problem and capture the flag. CTF is dead; Long live CTF.

9 Collaboration and Knowledge Transfer

Collaboration is a key, central aspect of all the ACTION Institute activities. Creating a vibrant exchange of ideas and influences among the Institute's organizations is paramount to the success of the Institute's mission. The transfer of knowledge and ideas to industry stakeholders, along with fostering collaborations among ACTION's member organizations, is listed as one of the Institute's three major goals (Section 1).

Year 3 demonstrated that ACTION's knowledge transfer agenda is maturing from aspiration to practice. TOUCAN, a large-scale dataset for training tool-using AI agents developed in collaboration with IBM, was released publicly on GitHub. CVE-GENIE, an automated multi-agent system for reproducing verifiable exploits, was released publicly. Team Shellphish's ARTIPHISHELL was released as open source in August

2025 alongside all other DARPA AIxCC finalists, with DARPA actively applying these cyber reasoning systems to open and closed-source repositories. Two startups founded by ACTION researchers, Virtue AI and Artiphishell, are commercializing ACTION research.

The plan for collaboration and knowledge transfer includes multiple steps to help the ACTION Institute faculty, research staff, and students as well as industrial partners identify the most appropriate paths and formats of collaboration. The plan includes the following components:

1. **Identify the key technologies for transfer.** Research, by nature, is exploratory and uncertain. As a result, it is not always possible to predict which research topic may generate marketable technologies. The ACTION Institute will organize an annual technology showcase and invite members of academia, industry, and government to evaluate the technologies that may have commercial value. This one-day Knowledge Transfer Expo will showcase a selection of the results of the Institute's research, with a focus on technology transfer and input on potential venues and domains for the real-world application of the Institute's novel ideas. The Institute will retain funds to be used for inviting members of industry, government, and academia to this event. Moreover, we will leverage our current industry collaborators, External Advisory Board, IBM, and collaborating researchers at DHS/CISA for updates and advice on key technologies and how to increase external participation in the Expo. ACTION faculty agreed to hold the annual Knowledge Transfer Expo at UC Santa Barbara each year, in conjunction with the yearly site visit.

2. **Protect intellectual property.** Marketable technologies may need to be protected as intellectual property. This may be at odds with the faculty of the ACTION Institute, who prefer to provide open access to their discoveries. The ACTION Institute will interact regularly with campus organizations such as UC Santa Barbara's Office of Technology and Industry Alliances (TIA) to obtain guidance on issues concerning IP. We will ask these offices to provide regular education sessions in the forms of webinars and online discussions to educate all researchers in ACTION about the differences between open access and protecting intellectual property. When necessary, the ACTION Institute will explore how to help and support inventors of marketable technologies, usually through the inventors' home institutions. The ACTION Institute will revisit its plans for Commercialization and Licensing (Section 12) on a regular basis, with the help of

campus technology specialists, and maintain compliance with the cooperative agreement between NSF that funds the AI Institute.

3. **Discover customers and their needs.** Academic research does not always produce marketable technologies. The NSF I-Corps program is designed to help researchers understand potential customers and their needs. The ACTION Institute will follow the best practices of I-Corps and, when relevant, encourage its researchers to discover customers and their needs. The researchers will learn how to fill the Business Model Canvas and conduct interviews to identify the market's needs.

4. **Develop strategies for transfers.** Different strategies will be appropriate for different technologies. For example, some technologies would be best licensed to well-established companies. Some other technologies are truly disruptive and should consider startups. Some technologies may be best transferred through open-source or non-profit organizations. By maintaining contact with campus technology specialists like UCSB's TIA, and through collaboration with various industry partners such as IBM, ACTION Institute will help researchers evaluate the best routes toward transferring their technologies. In Year 3, TOUCAN was transferred through open-source release in partnership with IBM. CVE-GENIE and ARTIPHISHELL followed similar open-source release strategies. These cases serve as models for Year 4 transfer decisions.

5. **Deepen government and agency engagement.** ACTION has established direct, ongoing relationships with federal agencies that go beyond conference presentations. Dr. Vigna briefed the NITRD Cyber Security and Information Assurance Interagency Working Group in July 2025. ACTION has been introduced to the Coalition for Secure AI (CoSAI), a cross-industry initiative advancing secure and trustworthy AI development, and plans to engage more actively with this community in Year 4. These relationships provide a direct channel for translating ACTION research into federal policy and operational practice.

6. **Create seed fund for knowledge transfer.** ACTION Institute will work with industrial partners, local and federal governments, incubators, venture studios, and investors to establish a seed fund that can help researchers move technology innovation from laboratories to markets by building minimum viable products. ACTION will continue to support researchers who identify opportunities for

spinout activity, building on the commercialization pathway demonstrated in Year 3.

10 Participation and Access Plan

The ACTION Institute is committed to broad participation and access in our research, as well as in our educational impact.

10.1 Vision

At the ACTION Institute, we believe broad participation is important for the creation of a global community that represents every voice, with the understanding that it is the harmony of these voices that fuels groundbreaking innovation. We are dedicated to not just including, but also uplifting individuals who have historically had limited access to opportunities in AI and computer security, recognizing that our collective strength can only be fully realized when everyone has an equal opportunity to contribute.

The ACTION Institute's goal is to drive forward the frontiers of AI and security by cultivating an open, accessible ecosystem where every member, be it students, faculty, or staff, can actualize their highest potential. Consistent with NSF's principle that the benefits of scientific discovery should create opportunities for all Americans everywhere, ACTION designs its activities to be open, non-exclusionary, and accessible to participants from a wide range of geographic, institutional, and socioeconomic backgrounds.

10.2 Definitions

Breadth of participation: Breadth encompasses cognitive diversity, including different ways of thinking, problem-solving, and approaching research and education, as well as the full range of backgrounds, experiences, geographic origins, and institutional affiliations that participants bring to the Institute.

Access: Access seeks to level the playing field by consciously addressing and removing structural hindrances that obstruct full participation and equal access. In the context of ACTION, this means ensuring that geography, institutional type, career stage, and socioeconomic background do not prevent qualified and motivated individuals from participating in and benefiting from the Institute's activities.

Inclusiveness: Being inclusive means going beyond numbers and quotas. It is about fostering a culture where everyone's voice is heard, and everyone feels they belong. It involves recognizing and valuing our inherent worth and dignity and making an active effort to solicit and consider the unique perspectives that each person brings.

10.3 Hiring Practices

Job Descriptions: When hiring institute staff and researchers, we will work closely with campus HR personnel to ensure that our job listings are free from biased language. Special attention will be given to not only what is said, but also how it is said. Minimum requirements will be considered so as to attract a wide array of candidates, and ensure the best of them are selected.

Hiring Panels: We will define clear rubrics so that candidates can be compared against one another fairly.

Shortlists: Reviews of candidate applications, job talks, and interviews will allow for anonymous feedback so that everyone can feel free to express their reservations.

On-boarding: Each new hire will be assigned a mentor from within the organization.

10.4 Decision-making and Governance

Leadership and Committees: Where possible, we will strive to keep committees and advisory boards and councils composed of a makeup that is broadly representative, reflecting a range of institutions, geographic regions, career stages, and areas of expertise. The goal is to ensure that a wide range of perspectives informs Institute leadership and decision-making.

Concerns about bias among leadership will be addressed by the External Advisory Board. Concerns about bias in the EAB and Student Advisory Council will be addressed by the Executive Committee. If necessary, relevant campus leadership, including ombudsmen, may be approached.

Strategic Planning: For any decision that significantly impacts the Institute, be it financial, operational, or otherwise, the decision-makers will strive to consider how outcomes might impact the most vulnerable members of the group.

10.5 Societal Impact

Ethical AI: Research initiatives will evaluate the societal impact, particularly ensuring AI is developed and applied in a way that is inclusive and equitable. Instances of bias found in training data will be disclosed immediately and addressed or mitigated.

Community Partnerships: The Institute will actively seek collaborations with organizations, government bodies, and other initiatives serving communities across the United States. We aim to apply AI for social good, in ways that are directly beneficial to these communities. EC members especially, but also Co-PIs, researchers, and staff, should announce their affiliation with ACTION when communicating and interacting with these organizations.

Scholarships and Internships: These activities will be set up with a specific focus on adhering to the broad strategy set forth in the above section on Broadening Participation. Internship and fellowship opportunities must be broadly advertised. When possible, internship organizers should ensure they do not overlook opportunities to advertise to organizations and groups who do not usually have access to the high-level research made possible by ACTION.

School Outreach: To ensure that more people are made aware of vocational and educational opportunities available to them, ACTION personnel will engage the K-12 community, starting with schools with close proximity to Institute campuses. As described in Section 6, ACTION will offer seminars and hacking competitions for both high-school and undergraduate students aimed at demystifying AI and promoting it as a career path. Our target for high school outreach should be at least 10 schools per year of ACTION operation, and possibly more.

10.6 Monitoring, Evaluation, and Assessment

The Institute's consideration of, and adherence to, this plan will be consistently monitored by both the ACTION Executive Committee and the Institute's External Evaluators. As discussed in the above section on Education and Workforce Development, the evaluation team has created a logic model for tracking progress against stated goals. The external evaluation team will conduct regular interviews with Institute personnel and participants to guarantee consistency between the above plan, what is reported by the Institute's Executive Committee, and what is perceived by the Institute's participants.

Progress toward achieving targets and goals will be included in quarterly and annual reports and presented to the External Advisory Board for feedback. Relevant campus leadership may receive updates as well, when requested. Deviations from the above plan must be explained and justified; any corrections must be made swiftly.

Consistent with NSF's updated guidance on broadening participation, any data collected or disaggregated as part of this monitoring effort will be conducted in a manner that is open, non-exclusionary, and focused on creating opportunities for all Americans. We will work with the evaluation team to design data collection instruments accordingly.

This plan may also be adjusted as needed, by recommendation of the Executive Committee, the External Advisory Board, the External Evaluators, and/or the NSF.

11 Management

Spanning 12 campuses, the ACTION Institute will be structured around four hubs. The main hub and lead of the ACTION Institute will be the University of California, Santa Barbara. The other major hubs are Purdue University, the University of Illinois Urbana-Champaign, and the Georgia Institute of Technology. The hubs implement a geographically distributed concentration of AI and security competence, each with slightly different foci and a combination of foundational and use-inspired research. The other institutions involved include Norfolk State University, Rutgers University, the University of Chicago, the University of Illinois at Chicago, the University of California, Berkeley, the University of Virginia, the University of Washington, and Western Washington University.

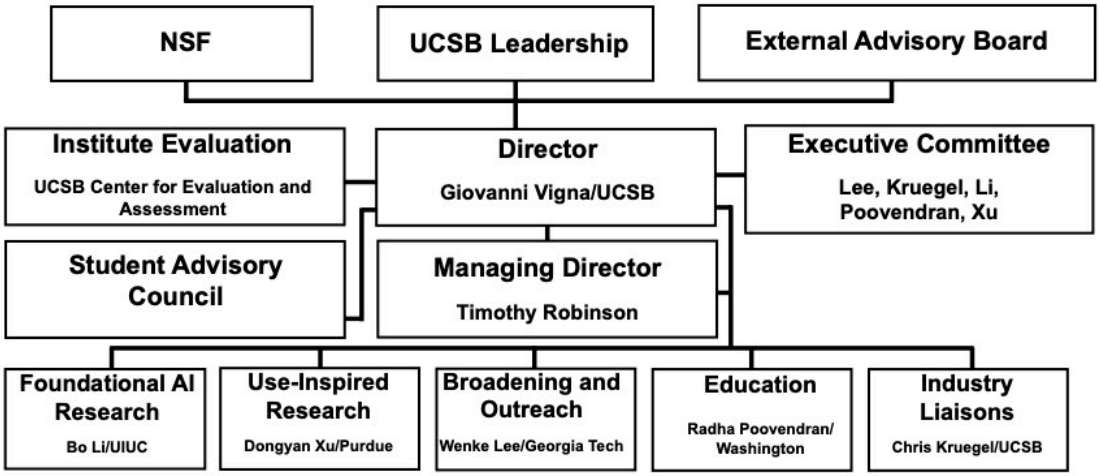


Figure 5: Overview of the ACTION Institute's management structure.

Figure 5 provides an overview of the ACTION Institute's management structure, and visualizes how the Institute will organize from bottom to top.

See Section 14 for a discussion of Continuity with respect to Institute management and decision making.

11.1 Organization

Director: At the center of the organization is Dr. Giovanni Vigna, the ACTION Institute PI and Director. Dr. Vigna has more than 25 years of experience in computer security and machine learning research and has published more than 200 papers on these subjects. He has extensive experience in managing multi-institution research efforts, and, in addition, he was the co-founder and CTO of Lastline, Inc., an AI-centered cybersecurity company, where he led the research team. The Director is responsible for all aspects of the Institute, including management, staffing, resource allocation, and communication with and reporting to the NSF. He is also the main point of contact for the Institute's External Evaluators, the External Advisory Board and Student Advisory Council, and primary campus (UCSB) leadership.

Managing Director: The Institute Director is supported by a Managing Director/Program Manager, hired through the lead ACTION campus (UCSB). The managing director is responsible for the daily management of the Institute as well as the

coordination among its members, the organization of events and meetings, the recruiting of talent (such as administrative personnel, students, and post-docs), the administration of funds, and the interaction with entities external to the Institute. The current managing director is Tim Robinson, a UCSB employee who has more than 15 years of experience with coordinating large research projects and centers, including collaborative efforts involving multiple campuses over multi-year periods of performance. Mr. Robinson has also helped establish connections between high school, undergraduate, and graduate training programs with a focus on broadening participation.

Executive Committee: Made up of five ACTION senior personnel, the Executive Committee (EC) is the primary advisory body to the Institute's Director, advising on matters related to the integration of activities across the institutions and research thrusts, progress in research projects, and the allocation of resources.

The EC assists the Director with decision making, planning, and an assurance that the goals of the Institute are being pursued. The three main goals of the ACTION Institute can be summarized as:

1. Conduct research into intelligent security agents capable of operating at speed and scale necessary to defend current and future threats to critical infrastructure;
2. Educate and train the next generation of workers who will operate at the intersection of AI and Security;
3. Transfer knowledge to industry and governmental stakeholders for further development.

Goal #1 can be further broken down into two sub-goals centered around fundamental research in AI (e.g., the development and deliverance of an AI stack) and use-inspired research on cybersecurity. Furthermore, the Institute's Broadening Participation and Outreach efforts, which are related to, and inform, our efforts in educational development, are significant enough to be considered on their own. To this end, the Institute's three main goals can be seen as the five programmatic pillars in Figure 5 upon which ACTION is built. Each member of the EC will assume responsibility for the leadership and management of one of the pillars. The Director has final authority over the assignment.

In Year 4 of ACTION, EC leadership roles will continue with five of the leading faculty researchers responsible for the initial ACTION program. Each of these five individuals has demonstrated expertise in their respective area and has indicated a willingness to

lead. The Foundational AI research lead is Dr. Bo Li, who has contributed work on integrating domain knowledge and logic reasoning abilities with statistical learning processes to enhance AI robustness against diverse attacks [115]. The Use-Inspired research lead is Dr. Dongyan Xu, who is the director of Purdue University's cybersecurity research and education center, CERIAS. Dr. Li and Dr. Xu have extensive experience in managing and coordinating large research programs and will continuously interact with each other to foster the virtuous cycle between foundational and use-inspired research activities discussed in Sections 2 through 4 of this document. The Knowledge Transfer and Industry Liaison lead is Dr. Christopher Kruegel, whose work in industry makes him well-qualified to coordinate outreach to industry and the industry collaboration program. Dr. Kruegel will act as a liaison between researchers, industry members, and technology transfer professionals at the various institutions. The Education lead, Dr. Radha Poovendran, will oversee the Institute's educational efforts. Dr. Poovendran has substantial experience in leading these types of activities and will work closely with education coordinators at partner institutions. Finally, the Broadening Participation and Outreach lead will coordinate activities related to diversity, inclusion, and impacts beyond the Institute's organizations. Dr. Wenke Lee will act as the lead for this pillar. He has extensive experience in conducting outreach and his proximity to the Atlanta public school system will become invaluable in future years.

EC members are expected to convene regularly, even if informally. The current meeting plan includes weekly virtual meetings for updates and task assignments. This meeting plan, as well as EC membership, shall be internally assessed annually for its efficacy and impact. Adjustments to the EC membership shall be considered as needed to assure smooth operations and a division of responsibility among the various goals and activities of the Institute. Dr. Vigna will consult with NSF (e.g., the Cognizant Program Officer) and other ACTION advisors, including the remaining members of the EC, about how to replace EC members. In the process of selecting an appropriate replacement for any absent EC members, he will consider representation from the various ACTION partnering campuses, along with expertise, availability, and willingness to lead the relevant focus area.

External Advisory Board: The ACTION Institute Director and EC will be supported by an External Advisory Board (EAB), composed of nationally and internationally recognized scientists and experienced industry leaders. The EAB's mission is to provide guidance, advice, and oversight for all of the Institute's activities, consistent with the vision, goals, and objectives described in this S&IP. The EAB is committed to helping the Institute advance science and research, with focus on the intersection of AI and cybersecurity, in an inclusive and ethical manner. The EAB reviews the Institute's

progress towards its goals, including research advancement, broadening participation, workforce education, and societal and technological impacts. The EAB provides internal advice and public accountability through annual meetings and quarterly check-ins.

The Institute intends to maintain an EAB of no less than five, and no more than nine members. As of this writing, the EAB consists of six members:

- Snehal Antani, CEO and Co-founder, Horizon3
- Mary Ann Blair, CISO, Carnegie Mellon University
- Tom Gillis, SVP and GM, Cisco Security Business Group
- Latanya Sweeney, Professor of the Practice of Government and Tech, Harvard
- Igor Tsyganskiy, CSO of Security, Microsoft
- Wen Masters, VP of Cyber Technologies, MITRE

The ACTION EC continues to propose, discuss, review, and invite new members, with particular consideration of leaders in academia who can contribute to the mission of the EAB.

An organizational charter, outlining in detail the formation, vision, and operational procedures for the EAB is on file with the Institute. This charter is revisited annually and amended as necessary.

To maintain the EAB, the EC will use their connections in both industry and academia to collect a diverse and cross-disciplinary Board who will have the necessary skills and experience to evaluate the ACTION Institute's results and planned activities, as well as the results of the external evaluation activities. The EC will propose and consider an initial list of candidates. They will invite a suitable subset to participate in the EAB, taking care to maintain a balance between members of industry and academia who have experience in computer security and/or AI. Even though the members of the Institute will maintain a constant dialogue with the members of the EAB, the EAB will review the Institute's activities once a year and provide a written report. With the help of ACTION's Managing Director, the EAB will be responsible for communicating this feedback to the members of the Institute; ACTION Leadership will include this assessment in their annual report.

Student Advisory Council: The ACTION Institute Director shall also be supported by a Student Advisory Council (SAC), which will be composed of student representatives from the various institutions and will be led by a Council head. In addition to providing to the Director the perspective of the students and postdocs involved in the research,

the SAC will promote student initiatives, support student mentoring and seminar activities, and help with the selection of ACTION interns. The initial SAC membership was selected from ACTION faculty nominations, and one of the SAC's first official responsibilities was to ratify an organizational charter. The charter, which provides details on the SAC operations, is kept on file with the Institute and made available to all members of the Council. Each year, the SAC is asked to revisit the charter and amend as needed.

National Science Foundation: The ACTION Institute has certain responsibilities to the National Science Foundation (NSF), as well as to the sponsoring partners: the Department of Homeland Security (DHS) and IBM. The responsibilities to NSF, which include submission of quarterly interim reports and an annual report, as well as an annual site visit, are outlined in the AI Institute's award documentation and cooperative agreement.

UCSB Leadership: The ACTION Institute also has reporting responsibilities to UCSB, the lead campus. UCSB's Vice Chancellor for Research, as well as the Dean of the College of Engineering and any participating Department Chairs, may also request updates on activity.

Institute Evaluation: For accountability, NSF also asks that the ACTION Institute retain the services of External Evaluators. The external evaluation team (UCSB Center for Evaluation & Assessment) leads this effort and helps with self-assessments and ensure that the Institute's stated goals are being met. The evaluation team will maintain and develop an evaluation plan and will be responsible for the steps needed to follow through with that plan. They will work with ACTION Leadership to collect the data necessary to determine progress against goals, including through surveys with ACTION participants and personnel. The evaluation team will include findings in an annual evaluation report, which will be provided to ACTION Leadership for assessment. This report will also be submitted to NSF as part of the Institute's annual report.

11.2 Research Prioritization and People Management

The ACTION Institute has structured its research plan (Sections 3 and 4) around eight interdependent research thrusts. Four thrusts concern the development of an AI stack for intelligent security agents; the other four explore deployment of that stack across the cyber-defense life-cycle. ACTION performers align themselves with those thrusts that resonate with their current research interests. Due to the integration of the research,

performers may participate in multiple thrusts and may find themselves adjusting and altering their focus and thrust concentrations.

Faculty associated with each thrust should meet regularly to provide updates with one another. The frequency of the meetings will be agreed upon by the faculty participants and enforced and arranged by an assigned thrust lead, who is responsible for the efforts of that thrust. The Institute's Managing Director may be enlisted for help with meeting logistics.

ACTION Thrust	**Thrust Lead**
AI-1: Learning & Reasoning w/ Domain Knowledge	Ambuj Singh
AI-2: Human-Agent Interaction	Ming Yin
AI-3: Multi-Agent Collaboration	Jie Gao
AI-4: Strategic Gaming & Tactical Planning	Joao Hespanha
SEC-1: Threat/Vulnerability Assessment	Gang Wang
SEC-2: Detection and Understanding	David Evans
SEC-3: Threat Intelligence	Nick Feamster
SEC-4: Response & Recovery	Elisa Bertino
Table 3: List of research thrusts faculty leads	

The assignment of thrust leads may change to accommodate expertise, future direction of the research, and availability and willingness to contribute to the research aims of ACTION. However, the Institute's research is first and foremost a collaborative effort. Thrust leads are a reflection of which ACTION performers are currently available to lead specific efforts. Thrust leaders will be regularly surveyed by the EC for their ability and willingness to contribute. Leadership responsibilities may rotate among performers as needed.

Although we believe each of the eight research thrusts serves a vital role in achieving the goals of ACTION, and each is important in its own right, the Institute will leverage the annual research meeting, the interaction with the External Advisory Board, and the feedback from the external evaluation team to define research priorities, and, when necessary, perform corrective actions to maintain an integrated and focused research program. These actions might reallocate funds to strengthen existing projects that are particularly promising or successful or be used to support exciting new directions.

Reallocation will be made by the Director in consultation with the EAB and the Executive Committee. After initial funding to establish ACTION's presence at each of the partnering campuses, which may be between one and three years depending on the direction of the Institute, internal proposals may be requested by ACTION Leadership to better focus the efforts of the Institute and the use of funds. Any thrust-related projects requiring a renewed allocation of funds will incur a review process, with review criteria consistent with the first two criteria (intellectual merit and broader impact) listed in the National Artificial Intelligence Research Institutes Program Solicitation (NSF 20-604). Review will be conducted by the EC and current thrust leads, with parties having a recognized conflict of interest recusing themselves from the decision. Details of the RFP process and review are discussed in the Section on Integration.

ACTION will promote a management culture that values inclusion and diverse viewpoints; the management goal is to prevent, reduce, and mitigate disputes among researchers. Best practices for communication, teamwork, and conflict resolution will be applied as part of a continuous process of improvement.

11.3 Key Personnel

The team involved in the ACTION Institute is composed of experts in Artificial Intelligence, Computer and Network Security, Machine Learning, Privacy, Game Theory, Data Science, Autonomous Agents, Malware Analysis, Threat Intelligence, and Social Networks, as shown in Table 4.

Participating personnel join one or more of the eight research thrusts according to their current research interests and ability to contribute to the tasks in progress. Monthly faculty meetings will allow cross-pollination of ideas. As the eight research thrusts are considerably integrated and dependent on one another, ACTION personnel should discover new collaborations with one another through their ACTION connections.

Faculty currently unaffiliated with ACTION are welcome to join the cohort, pending approval by the EC. The involvement of additional campuses would require subaward agreements at the campus level and cannot be made without the approval of UC Santa Barbara and NSF.

Name	**Site**	**Expertise**	**Thrust**

Giovanni Vigna	UCSB, PI	Network Security, ML, Threat Intelligence	AI-4, SEC-1, SEC-2, SEC-3, SEC-4
Dongyan Xu	Purdue, Co-PI	Systems Security, Attack Data Analytics	SEC-1, SEC-3, SEC-4
Dawn Song	Berkeley, Co-PI	AI, Vulnerability Analysis, Threat Modeling	AI-1, SEC-1, SEC-2, SEC-3
Wenke Lee	GIT, Co-PI	Malware Analysis, Network Security, ML	AI-4, SEC-1, SEC-3, SEC-4
David Evans	Virginia	Trustworthy ML, Privacy, Software Security	SEC-1, SEC-2, SEC-3
Stuart Russell	Berkeley	AI, ML, Reasoning	AI-1
Christopher Kruegel	UCSB	Network Analysis, ML, Cloud Security	SEC-1, SEC-2, SEC-3
Bo Li	UIUC	AI, Reasoning, Game Theory, Network Security	AI-1, AI-4, SEC-3, SEC-4
Joao Hespanha	UCSB	Game Theory, Autonomous Agents, Controls	AI-3, AI-4, SEC-2
Xiaoming Huo	GIT	Data Science, ML, Multi-Agent Collaboration	AI-3
Pieter Abbeel	Berkeley	AI, Robotics, Reinforcement Learning	AI-2, AI-3, SEC-1, SEC-2
Ninghui Li	Purdue	Differential Privacy, Reasoning, Adversarial ML	AI-1, SEC-1, SEC-2
Elisa Bertino	Purdue	Wireless Network Security, Transfer Learning	SEC-2, SEC-4
David Wagner	Berkeley	Adversarial ML, Privacy, Software Security	SEC-2, SEC-4
Bin Yu	Berkeley	Trustworthy AI, ML, Statistics, Data Science	AI-1
Z. Berkay Celik	Purdue	ML, Attack Analysis and Attribution	SEC-1, SEC-2, GATE Lead
Tonya Fields	Norfolk	ML, Incident Response, Malware Analysis	SEC-3
Ming Yin	Purdue	Human-Agent Interaction, AI, Crowd Intelligence	AI-2
Ambuj Singh	UCSB	AI, ML, Data Science, Social	AI-1, AI-2, AI-3

		Networks	
Yung-Hsiang Lu	Purdue	Low-Power Computing, ML, Computer Vision	SEC-3
Jie Gao	Rutgers	Social Networks, Data Privacy, ML	AI-3
Gang Wang	UIUC	Network Security, ML, Data Mining	SEC-1, SEC-2, SEC-4
Bing Liu	UI Chicago	Lifelong Learning, Open-world ML, AI	AI-1, AI-2
Nick Feamster	UChicago	ML, Network Security, Privacy	SEC-2, SEC-3
Radha Poovendran	Washington	Threat Modeling, Multiagent Learning	AI-2, AI-3, AI-4, SEC-2, SEC-3, SEC-4
Sewoong Oh	Washington	Theoretical ML, Networks, Differential Privacy	SEC-2
Mary Ann Hoppa	Norfolk	Cybersecurity, Cyberpsychology	AI-2

Table 4: ACTION key personnel

12 Intellectual Property

The ACTION Institute is committed to the principles of openness, transparency, and the widest possible dissemination of research findings. The Institute recognizes that a judicious intellectual property (IP) strategy is crucial to meeting these goals, while also providing opportunities for commercialization and practical implementation of research findings. The following document outlines our policy regarding IP processes and the management of data (both research data and evaluation data).

12.1 Publications, Datasets, and Code

The ACTION Institute encourages researchers to share their work as pre-prints before undergoing peer review. This ensures immediate open access to research outputs. All publications resulting from Institute-funded projects should be published in open-access journals or conferences or made open-access within a reasonable period after publication. In addition, wherever possible and ethical, datasets should be made publicly accessible, and software code developed under Institute-funded projects should be shared under open-source licenses whenever feasible.

12.2 Commercialization and Licensing

Wherever feasible, IP will be shared under permissive open-source licenses such as BSD, Apache, or MIT. In instances where open-source dissemination is impractical, the Institute may pursue traditional patent protection to encourage the commercialization of technologies. Specific arrangements for IP rights in third-party collaborations will be outlined in separate agreements but will be influenced by the Institute's commitment to openness and dissemination and in congruence with the award terms from the NSF. Finally, all third-party collaborations must be transparent and adhere to the Institute's openness policies.

12.3 Research Data

Types of Data. This project will generate data designed to study the effectiveness of using intelligent agents and other AI-based capabilities to improve the security of computer systems. To this end, the data that will be collected will include data associated with the intelligent agents (e.g., knowledge bases) as well as logs and traces associated with the components of networked systems, such as system call sequences, network dumps, and API logs. In addition, the research carried out at the ACTION Institute will include data about software components and the associated vulnerabilities, as well as information about the tools used by attackers (e.g., malware samples and attack tools).

Standards. Data will be collected using state-of-the-art formats, with a preference for JSON to represent metadata (e.g., the meta-information about a network dump, which might include the location where the network traffic was collected or specific filters that were applied to filter traffic). All datasets will be required to store a description of the data in a readme.txt file, accompanying the data. Documentation will be stored in ASCII, Markdown, and PDF formats.

Access and Sharing Policies. The research data from this project will be stored using standard, well-accepted techniques and procedures to ensure that the research community has long-term access to the data. The project team will create a dedicated website that will index the various data elements and put them in relation to the Institute's research thrusts.

The principal investigators on the project and their institutions will hold the copyright for the research data they generate but will be encouraged to grant redistribution rights for purposes of data sharing.

Data Archives. In addition to distributing the data from a project website, future long-term use of the data will be ensured by using suitable long-term approaches, which might be different depending on the type and size of the data, ensuring that best practices in digital preservation will safeguard the files.

To guarantee availability and resiliency against failures, corruption, and attacks, the data for the project will be stored in multiple copies, using various backup techniques, including continuous mirroring and versioned backups. The various copies will be geographically distributed.

12.4 Evaluation Data

The data management plan associated with the evaluation of the ACTION Institute is designed to be consistent with the NSF Policy on Dissemination and Sharing of Research Results (AAG Chapter VI.D.4). In accordance with this policy, this plan does not include preliminary analyses (including raw data), drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues. Further, data to enable peer review and publication/dissemination and/or to protect intellectual property may be temporarily withheld from distribution and other proposed data management.

Data Description. The proposed study will involve collecting two categories of data: primarily, student data (high school, undergraduate, and graduate program evaluation data) and, secondarily, administrative data collected as part of the participation in program activities. Both categories of data might include qualitative interviews, document reviews and analysis, and quantitative survey and administrative data. All procedures required by each institution's Institutional Review Board for the management and storage of data collected for this research will be followed.

Responsibility. The University of California, Santa Barbara is responsible for all data-related materials and oversight of data collection, secure storage, and analysis for the study. Graduate students and computer services staff, under the direction of the external evaluation team, will conduct database management and maintenance. All data will be securely archived by UCSB, and de-identified and aggregated before being provided to the program director to ensure the participants' confidentiality.

Designated Archive. All data will be securely archived by UCSB, and de-identified and aggregated before being provided to the program director to ensure the participants' confidentiality.

Access and Sharing. The data resulting from the program evaluation component of this project will be made available for use as soon as completed and no later than the conclusion of the project. These materials will be available as standard Microsoft Office files, PDF documents, and tab-delimited files. Any data shared would be de-identified to ensure the confidentiality of participants. The data resulting from the administrative research component of this project, including digitized paper administrative records, interview transcripts in raw form, and individual-level administrative data will only be accessible to IRB-approved study teams to preserve the confidentiality of participants. However, summary data, including detailed coding guides, will be produced and shared through available, relevant research networks.

Selection and Retention. The present study will collect and track data on students who participate in the various educational programs offered by the ACTION Institute.

13 Ethics

The ACTION Institute aims to nucleate and accelerate innovation through research collaborations, but also engage all participants and the broader community (including academic, industrial, governmental, and non-profit parties) in discussing the meaning of the Institute's discoveries and amplifying their significance. All partner universities engaged in research and collaborators are expected to conduct their work with integrity and intellectual honesty at all times, consistently with the ethical code of conduct that all our partner universities embrace. Mentors, in particular, are expected to foster these same high standards in their students, modeling good conduct and encouraging further training in professional and responsible conduct of research (RCR).

Ethics Training. All partner Institutions, except NSU, subscribe to the CITI program, which is part of the mandatory ethics training for students, fellows, and postdoctoral researchers. NSU provides in-person training on RCR that is consistent with NSF requirements. In addition to standard training modules required by all our institutions, which cover discipline-specific topics such as research misconduct, data management, authorship, peer review, conflict of interest, mentoring, and collaborative research, the ACTION Institute will also sponsor a series of webinars (three per year) in Year 4 on topics such as: "What is the ethics of AI?"; "What ethical issues arise from AI-powered

cybersecurity?"; "What are the ethics of counter-cyber-attacks?"; "How do we keep AI safe from adversaries?"; "Personal responsibility in the age of AI". The core partner institutions will take turns hosting the seminars, assigning a graduate student the role of host and discussion moderator, thus providing a further opportunity for leadership development. In selecting topics and speakers, the ACTION Institute will leverage existing resources and expertise at the partner campuses. At UCSB, the Center for Information Technology and Society seeks to understand and foster dialogue around the use and social effects of information technologies. The ACTION Institute faculty, postdoctoral researchers, and graduate and undergraduate students will be required to attend at least two lectures each year. The ACTION Institute Managing Director will monitor attendance. All lectures will be made available through Zoom, and will be recorded and posted on the institute website.

Authorship. We are committed to ensuring that all investigators, particularly students, are credited with their contributions to the joint research program. As has been our practice, publication authorship will be based on the relative scientific contributions and patent inventorship on US patent law; authorship will be discussed by project teams as each study is initiated and updated as necessary. Final decisions on authorship will be the responsibility of the involved investigators, with disputes resolved by the Executive Committee.

Ethical Responsibilities in Agentic AI Research. ACTION's research operates at the intersection of offensive and defensive cybersecurity, and the institute takes its ethical responsibilities in this space seriously. Several research outputs, including CVE-GENIE, which automates exploit reproduction, and ARTIPHISHELL, an autonomous cyber reasoning system, represent dual-use technologies with clear potential for misuse. In each case, release decisions are made deliberately, with attention to responsible disclosure norms and the defensive value of making these tools available to the security research community. As ACTION's research agenda advances in Year 4, new tools and agents with dual-use potential will continue to emerge, and the institute is committed to applying the same deliberate approach to each release decision.

For agentic AI systems deployed in or connected to real infrastructure, ACTION researchers apply standard security research ethics principles: tools are evaluated in controlled environments before release, and datasets such as MABEL and TOUCAN are constructed from publicly available or synthetically generated data to avoid privacy implications. As agents become more capable of taking consequential actions with limited human oversight, questions of accountability, explainability, and controllability take on increasing ethical weight. ACTION's research explicitly addresses these

concerns through the human-in-the-loop paradigms developed in Thrust AI-2, the agent hardening work in Thrust SEC-4, and the misuse mitigation requirements described in Research Risk 4 in Section 5. The principle that certain agent actions should always require human approval is not merely an operational constraint but an ethical one, and ACTION researchers are expected to design their systems accordingly.

The evaluation activities described in Section 6.7, including surveys and interviews with summer research participants and competition participants, are conducted under appropriate human subjects review protocols at the relevant institutions.

Finally, ACTION is developing ethics training content appropriate to the AI and cybersecurity intersection, an area not well served by general AI ethics curricula. In Year 3, ACTION leadership evaluated a Coursera course in AI ethics developed by the Polytechnic University of Milan, recommended by the Special Competitive Studies Project. A decision about its incorporation into institute requirements will be made during Year 4, and ACTION will report on progress in the next annual report. More broadly, the quarterly ethics webinar series described above provides an ongoing forum for institute participants to engage with the specific ethical challenges that arise at the intersection of AI and cybersecurity.

Responsible Disclosure. At the ACTION Institute, responsible disclosure of security vulnerabilities is paramount and is conducted in accordance with established industry standards, such as the ISO/IEC 29147 guidelines on vulnerability disclosure. When a security vulnerability is identified in any software, hardware, or system, either through our research activities or by external parties, the initial steps involve internal verification and risk assessment to ascertain the severity and potential impact of the vulnerability. The team then promptly notifies the vendor or entity responsible for the affected technology, providing them with detailed information and suggested mitigation measures. This is done confidentially to allow the affected party an adequate window of time, typically between 45 to 90 days, to address the issue before public disclosure. During this period, the Institute maintains open channels of communication with the concerned parties to provide additional guidance or clarification. If the vulnerability is not addressed within the agreed timeframe, the Institute may escalate the issue to include governmental bodies or other overseeing agencies to ensure the vulnerability is taken seriously and remedied in a timely manner. Following successful mitigation, the Institute aims to produce a comprehensive public advisory, in collaboration with the affected party, that includes technical details, risk assessments, and recommended actions for stakeholders. This advisory is then disseminated through trusted channels and may be accompanied by academic publications or white papers,

all in line with our commitment to open dissemination of knowledge and ensuring cybersecurity for the broader community.

14 Continuity

Ensuring the continuous operation of the ACTION Institute is crucial for sustaining its mission and ongoing programs. We strive to mitigate disruptions caused by the absence or loss of key personnel.

Institute records are maintained under UCSB's enterprise Google agreement, with access arranged for successors. The UC Santa Barbara enterprise agreement provides a solution to long-term storage for the sake of continuity. All documents can be shared with key personnel as deemed appropriate, such as for succession. ACTION also maintains a regularly-used and updated YouTube channel.

In case of loss of key personnel (e.g., Director, Managing Director, Executive Committee members), we will first determine the impact of the absence and identify any immediate operational bottlenecks. Until a long-term plan is set, daily responsibilities will be transferred to a pre-identified researcher or staff member familiar with the role. In addition, we will immediately notify relevant stakeholders and team members about the temporary changes. Table 5 shows key personnel and their interim counterparts.

In case of a long-term loss of a key role, we will ensure that any ongoing contracts, research commitments, or financial transactions are reviewed and maintained. We will also conduct an in-depth review of affected operations and timelines for critical projects. Finally, if the absence is prolonged or permanent, we will confirm a successor or initiate a recruitment process. If necessary, UCSB's Vice Chancellor for Research, in consultation with NSF, will appoint an interim director and form a search committee of faculty and administrators at UCSB for a replacement, internally or through a national search.

Key Personnel	**Interim Substitutes**
Giovanni Vigna (Director)	Christopher Kruegel
Timothy Robinson (Managing Director)	Giovanni Vigna
Bo Li (AI Research Lead)	Xiaoming Huang or Ambuj Singh
Dongyan Xu (Security Research Lead)	Wenke Lee

Wenke Lee (Broadening and Outreach)	Jie Gao
Radha Poovendran (Education Lead)	Gang Wang
Christopher Kruegel (Industry Lead)	Yung-Hsiang Lu

Table 5: List of key personnel and their interim substitutes.

We will follow a similar process for the permanent replacement of other key personnel (e.g., Co-PIs and research leads). ACTION is committed to maintaining a presence at each of the partnering institutions and will work with a partnering campus's leadership, offices of academic personnel, and/or human resources to fill personnel vacancies.

By adhering to this continuity plan, we aim to minimize the impact of any unplanned disruptions and ensure the ongoing success and resilience of the institute.

15 References Cited

[1] Mandiant, "Highly evasive attacker leverages SolarWinds supply chain to compromise multiple global victims with SUNBURST backdoor," Dec. 2020. [Online]. Available: https://www.mandiant.com/resources/evasive-attacker-leverages-solarwinds-supply-chain-compromises-with-sunburst-backdoor

[2] LunaSec, "Log4Shell: RCE 0-day exploit found in log4j 2, a popular Java logging package," Dec. 2021. [Online]. Available: https://www.lunasec.io/docs/blog/log4j-zero-day/

[3] J. Mirkovic, D. Balenson, and B. Kocoloski, "Enabling reproducibility through the SPHERE research infrastructure," 2024. [Online]. Available: https://sphere-project.net/

[4] B. Yu and R. L. Barter, *Veridical Data Science: The Practice of Responsible Data Analysis and Decision Making*. Cambridge, MA, USA: MIT Press, 2024. [Online]. Available: https://vdsbook.com

[5] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Hum. Factors*, vol. 46, no. 1, pp. 50–80, 2004.

[6] Z. Buçinca, M. B. Malaya, and K. Z. Gajos, "To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making," *Proc. ACM Hum.-Comput. Interact.*, vol. 5, no. CSCW1, pp. 1–21, 2021.

[7] C.-W. Chiang and M. Yin, "You'd better stop! Understanding human reliance on machine learning models under covariate shift," in *Proc. 13th ACM Web Sci. Conf.*, 2021, pp. 120–129.

[8] Z. Ashktorab *et al.*, "AI-assisted human labeling: Batching for efficiency without overreliance," *Proc. ACM Hum.-Comput. Interact.*, vol. 5, no. CSCW1, pp. 1–27, 2021.

[9] P. J. Aubusson, A. G. Harrison, and S. M. Ritchie, "Metaphor and analogy," in *Metaphor and Analogy in Science Education*. Dordrecht, The Netherlands: Springer, 2006, pp. 1–9.

[10] S. Vosniadou and A. Ortony, *Similarity and Analogical Reasoning*. Cambridge, U.K.: Cambridge Univ. Press, 1989.

[11] A. Xiong, R. W. Proctor, W. Yang, and N. Li, "Embedding training within warnings improves skills of identifying phishing webpages," *Hum. Factors*, vol. 61, no. 4, 2019, doi: 10.1177/0018720818810942.

[12] D. Long and B. Magerko, "What is AI literacy? Competencies and design considerations," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2020, pp. 1–16.

[13] C.-W. Chiang and M. Yin, "Exploring the effects of machine learning literacy interventions on laypeople's reliance on machine learning models," in *Proc. 27th Int. Conf. Intell. User Interfaces*, 2022.

[14] G. Bansal, B. Nushi, E. Kamar, W. S. Lasecki, D. S. Weld, and E. Horvitz, "Beyond accuracy: The role of mental models in human-AI team performance," in *Proc. AAAI Conf. Hum. Comput. Crowdsourcing*, vol. 7, no. 1, 2019, pp. 2–11.

[15] X. Wang and M. Yin, "Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making," in *Proc. 26th Int. Conf. Intell. User Interfaces*, 2021, pp. 318–328.

[16] G. Bansal, B. Nushi, E. Kamar, D. S. Weld, W. S. Lasecki, and E. Horvitz, "Updates in human-AI teams: Understanding and addressing the performance/compatibility tradeoff," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, 2019, pp. 2429–2437.

[17] S. Plous, *The Psychology of Judgment and Decision Making*. New York, NY, USA: McGraw-Hill, 1993.

[18] A. Tversky and D. Kahneman, "The framing of decisions and the psychology of choice," in *Behavioral Decision Making*. Boston, MA, USA: Springer, 1985, pp. 25–41.

[19] A. Tversky and D. Kahneman, "Advances in prospect theory: Cumulative representation of uncertainty," *J. Risk Uncertainty*, vol. 5, no. 4, pp. 297–323, 1992.

[20] D. Kahneman and A. Tversky, "Prospect theory: An analysis of decision under risk," in *Handbook of the Fundamentals of Financial Decision Making: Part I*. Singapore: World Scientific, 2013, pp. 99–127.

[21] C. Gonzalez, J. F. Lerch, and C. Lebiere, "Instance-based learning in dynamic decision making," *Cogn. Sci.*, vol. 27, no. 4, pp. 591–635, 2003.

[22] X. Wang, Z. Lu, and M. Yin, "Will you accept the AI recommendation? Predicting human behavior in AI-assisted decision making," in *Proc. ACM Web Conf.*, 2022.

[23] W. Ye, F. Bullo, N. Friedkin, and A. K. Singh, "Modeling human-AI team decision making," 2022, *arXiv:2201.02759*.

[24] G. Bansal, B. Nushi, E. Kamar, E. Horvitz, and D. S. Weld, "Is the most accurate AI the best teammate? Optimizing AI for teamwork," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 13, 2021, pp. 11405–11414.

[25] V. D. Veksler and N. Buchler, "Know your enemy: Applying cognitive modeling in security domain," in *Proc. Annu. Meeting Cogn. Sci. Soc. (CogSci)*, 2016.

[26] V. D. Veksler, N. Buchler, C. G. LaFleur, M. S. Yu, C. Lebiere, and C. Gonzalez, "Cognitive models in cybersecurity: Learning from expert analysts and predicting attacker behavior," *Front. Psychol.*, vol. 11, Art. no. 1049, 2020.

[27] V. D. Veksler, N. Buchler, B. E. Hoffman, D. N. Cassenti, C. Sample, and S. Sugrim, "Simulations in cyber-security: A review of cognitive modeling of network attackers, defenders, and users," *Front. Psychol.*, vol. 9, Art. no. 691, 2018.

[28] K. Veeramachaneni, I. Arnaldo, V. Korrapati, C. Bassias, and K. Li, "AI^2: Training a big data machine to defend," in *Proc. IEEE 2nd Int. Conf. Big Data Secur. Cloud (BigDataSecurity)*, 2016, pp. 49–54.

[29] S. Das, W.-K. Wong, T. Dietterich, A. Fern, and A. Emmott, "Incorporating expert feedback into active anomaly discovery," in *Proc. IEEE 16th Int. Conf. Data Mining (ICDM)*, 2016, pp. 853–858.

[30] S. Russell, "Unifying logic and probability," *Commun. ACM*, vol. 58, no. 7, pp. 88–97, 2015.

[31] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Definitions, methods, and applications in interpretable machine learning," *Proc. Nat. Acad. Sci.*, vol. 116, no. 44, pp. 22071–22080, 2019.

[32] B. Xiao, Q. Lu, B. Ramasubramanian, A. Clark, L. Bushnell, and R. Poovendran, "FRESH: Interactive reward shaping in high-dimensional state spaces using human feedback," in *Proc. 19th Int. Conf. Auton. Agents Multiagent Syst. (AAMAS)*, 2020, pp. 1512–1520.

[33] L. Guan, M. Verma, S. Guo, R. Zhang, and S. Kambhampati, "Widening the pipeline in human-guided reinforcement learning with explanation and context-aware data augmentation," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 34, 2021.

[34] J. Kim, C. J. Banks, and J. A. Shah, "Collaborative planning with encoding of users' high-level strategies," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017.

[35] E. D. Gennatas *et al.*, "Expert-augmented machine learning," *Proc. Nat. Acad. Sci.*, vol. 117, no. 9, pp. 4571–4577, 2020.

[36] C. Heinze-Deml, M. H. Maathuis, and N. Meinshausen, "Causal structure learning," *Annu. Rev. Statist. Appl.*, vol. 5, pp. 371–391, 2018.

[37] J.-P. Pellet and A. Elisseeff, "Using Markov blankets for causal structure learning," *J. Mach. Learn. Res.*, vol. 9, no. 7, 2008.

[38] J. Langford and T. Zhang, "The epoch-greedy algorithm for contextual multi-armed bandits," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 20, 2007.

[39] A. Swaminathan and T. Joachims, "Counterfactual risk minimization: Learning from logged bandit feedback," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 814–823.

[40] T. Chakraborti, S. Sreedharan, and S. Kambhampati, "Human-aware planning revisited: A tale of three models," in *Proc. IJCAI-ECAI XAI/ICAPS XAIP Workshops*, 2018.

[41] T. Chakraborti, S. Sreedharan, and S. Kambhampati, "The emerging landscape of explainable automated planning and decision making," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2020, pp. 4803–4811.

[42] T. O'Neill, N. McNeese, A. Barron, and B. Schelble, "Human–autonomy teaming: A review and analysis of the empirical literature," *Hum. Factors*, 2020, doi: 10.1177/0018720820960865.

[43] G. Ghasemiesfeh, R. Ebrahimi, and J. Gao, "Complex contagion and the weakness of long ties in social networks: Revisited," in *Proc. 14th ACM Conf. Electron. Commerce (EC)*, 2013, pp. 507–524, doi: 10.1145/2482540.2482550.

[44] J. Gao, B. Li, G. Schoenebeck, and F.-Y. Yu, "Engineering agreement: The naming game with asymmetric and heterogeneous agents," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 537–543.

[45] R. Ebrahimi, J. Gao, G. Ghasemiesfeh, and G. Schoenebeck, "Complex contagions in preferential attachment models and other time-evolving networks," *IEEE Trans. Netw. Sci. Eng.*, vol. 4, no. 4, pp. 201–214, 2017.

[46] J. Gao, G. Schoenebeck, and F.-Y. Yu, "The volatility of weak ties: Co-evolution of selection and influence in social networks," in *Proc. 18th Int. Conf. Auton. Agents Multiagent Syst. (AAMAS)*, May 2019, pp. 619–627.

[47] J. Gao, G. Ghasemiesfeh, G. Schoenebeck, and F.-Y. Yu, "General threshold model for social cascades: Analysis and simulations," in *Proc. ACM Conf. Econ. Comput. (EC)*, 2016, pp. 617–634, doi: 10.1145/2940716.2940778.

[48] H. Wang, F. Luo, and J. Gao, "Co-evolution of opinion and social tie dynamics towards structural balance," in *Proc. ACM-SIAM Symp. Discrete Algorithms (SODA)*, Jan. 2022.

[49] N. Li, W. H. Winsborough, and J. C. Mitchell, "Distributed credential chain discovery in trust management," *J. Comput. Secur.*, vol. 11, no. 1, pp. 35–86, 2003.

[50] H.-P. Lam, G. Governatori, K. Satoh, and H. Hosobe, "Distributed defeasible speculative reasoning in ambient environment," in *Proc. Int. Workshop Comput. Log. Multi-Agent Syst. (CLIMA)*, 2012, pp. 43–60.

[51] C. Baral, M. Gelfond, and J. N. Rushton, "Probabilistic reasoning with answer sets," *Theory Pract. Log. Program.*, vol. 9, no. 1, pp. 57–144, 2009, doi: 10.1017/S1471068408003645.

[52] W. Brendel, J. Rauber, and M. Bethge, "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018.

[53] A. Bahramali, M. Nasr, A. Houmansadr, D. Goeckel, and D. Towsley, "Robust adversarial attacks against DNN-based wireless communications systems," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur. (CCS)*, 2021, pp. 126–140.

[54] L. Chen *et al.*, "Decision transformer: Reinforcement learning via sequence modeling," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 34, 2021.

[55] B. Xiao, B. Ramasubramanian, and R. Poovendran, "Agent-temporal attention for reward redistribution in episodic multi-agent reinforcement learning," in *Proc. Int. Conf. Auton. Agents Multiagent Syst. (AAMAS)*, 2022.

[56] MITRE, "MITRE ATT&CK," 2021. [Online]. Available: https://attack.mitre.org

[57] Y. Ji *et al.*, "RAIN: Refinable attack investigation with on-demand inter-process information flow tracking," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur. (CCS)*, 2017, pp. 377–390.

[58] A. Alsaheel *et al.*, "ATLAS: A sequence-based learning approach for attack investigation," in *Proc. 30th USENIX Secur. Symp.*, Aug. 2021.

[59] S. Moothedath, D. Sahabandu, A. Clark, S. Lee, W. Lee, and R. Poovendran, "Multi-stage dynamic information flow tracking game," in *Proc. Int. Conf. Decis. Game Theory Secur. (GameSec)*, 2018, pp. 80–101.

[60] S. Moothedath *et al.*, "A game-theoretic approach for dynamic information flow tracking to detect multistage advanced persistent threats," *IEEE Trans. Autom. Control*, vol. 65, no. 12, pp. 5248–5263, 2020.

[61] D. Sahabandu, J. Allen, S. Moothedath, L. Bushnell, W. Lee, and R. Poovendran, "Quickest detection of advanced persistent threats: A semi-Markov game approach," in *Proc. ACM/IEEE 11th Int. Conf. Cyber-Phys. Syst. (ICCPS)*, 2020, pp. 9–19.

[62] D. Sahabandu, S. Moothedath, J. Allen, L. Bushnell, W. Lee, and R. Poovendran, "A reinforcement learning approach for dynamic information flow tracking games for detecting advanced persistent threats," 2020, *arXiv:2007.00076*.

[63] D. Sahabandu, B. Xiao, A. Clark, S. Lee, W. Lee, and R. Poovendran, "DIFT games: Dynamic information flow tracking games for advanced persistent threats," in *Proc. IEEE Conf. Decis. Control (CDC)*, 2018, pp. 1136–1143.

[64] M. Zinkevich, M. Johanson, M. Bowling, and C. Piccione, "Regret minimization in games with incomplete information," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 20, 2007.

[65] B. Li and Y. Vorobeychik, "Scalable optimization of randomized operational decisions in adversarial classification settings," in *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*, 2015, pp. 599–607.

[66] A. Clark, K. Sun, L. Bushnell, and R. Poovendran, "A game-theoretic approach to IP address randomization in decoy-based cyber defense," in *Proc. Int. Conf. Decis. Game Theory Secur. (GameSec)*, 2015, pp. 3–21.

[67] S. Jajodia, A. K. Ghosh, V. Swarup, C. Wang, and X. S. Wang, Eds., *Moving Target Defense: Creating Asymmetric Uncertainty for Cyber Threats* (Advances in Information Security, vol. 54). New York, NY, USA: Springer, 2011.

[68] B. Wang *et al.*, "Neural Cleanse: Identifying and mitigating backdoor attacks in neural networks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, 2019, pp. 707–723.

[69] A. Rajabi, B. Ramasubramanian, A. A. Maruf, and R. Poovendran, "Privacy-preserving reinforcement learning beyond expectation," 2022, *arXiv:2203.10165*.

[70] Z. Lu, C. Wang, and S. Zhao, "Cyber deception for computer and network security: Survey and challenges," 2020, *arXiv:2007.14497*.

[71] Z. Chen and B. Liu, *Lifelong Machine Learning*. San Rafael, CA, USA: Morgan & Claypool, 2018.

[72] B. Liu, S. Mazumder, E. Robertson, and S. Grigsby, "AI autonomy: Self-initiation, adaptation and continual learning," 2022, *arXiv:2203.08994*.

[73] E. Delage and Y. Ye, "Distributionally robust optimization under moment uncertainty with application to data-driven problems," *Oper. Res.*, vol. 58, no. 3, pp. 595–612, 2010.

[74] X. Liu, W. Kong, S. Kakade, and S. Oh, "Robust and differentially private mean estimation," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 34, 2021, pp. 3887–3901.

[75] X. Liu, W. Kong, and S. Oh, "Differential privacy and robust statistics in high dimensions," 2021, *arXiv:2111.06578*.

[76] F. Valeur, G. Vigna, C. Kruegel, and R. A. Kemmerer, "A comprehensive approach to intrusion detection alert correlation," *IEEE Trans. Dependable Secure Comput.*, vol. 1, no. 3, pp. 146–169, Jul.–Sep. 2004.

[77] A. S. Li, A. Iyengar, A. Kundu, and E. Bertino, "Revisiting concept drift in Windows malware detection: Adaptation to real drifted malware with minimal samples," in *Proc. Netw. Distrib. Syst. Secur. Symp. (NDSS)*, 2025.

[78] A. N. Angelopoulos, S. Bates, C. Fannjiang, M. I. Jordan, and T. Zrnic, "Prediction-powered inference," *Science*, vol. 382, no. 6671, pp. 669–674, 2023.

[79] X. Jiang *et al.*, "NetDiffusion: Network data augmentation through protocol-constrained traffic generation," *Proc. ACM Meas. Anal. Comput. Syst.*, 2024.

[80] A. Chu *et al.*, "NetSSM: Multi-flow and state-aware network trace generation using state-space models," *Proc. ACM Netw.*, vol. 4, no. CoNEXT1, pp. 1–24, 2026.

[81] G. Wan, S. Liu, F. Bronzino, N. Feamster, and Z. Durumeric, "CATO: End-to-end optimization of ML-based traffic analysis pipelines," in *Proc. 22nd USENIX Symp. Netw. Syst. Des. Implement. (NSDI)*, 2025.

[82] Y. Shoshitaishvili *et al.*, "Mechanical Phish: Resilient autonomous hacking," *IEEE Secur. Privacy*, vol. 16, no. 2, Mar. 2018.

[83] Center for AI Safety, "CAIS compute cluster," 2024. [Online]. Available: https://www.safe.ai/work/compute-cluster

[84] U.S. Bureau of Labor Statistics, "Software developers, quality assurance analysts, and testers," in *Occupational Outlook Handbook*, 2023. [Online]. Available: https://www.bls.gov/ooh/computer-and-information-technology/software-developers.htm

[85] U.S. Bureau of Labor Statistics, "Information security analysts," in *Occupational Outlook Handbook*, 2023. [Online]. Available: https://www.bls.gov/ooh/computer-and-information-technology/information-security-analysts.htm

[86] Carnegie Mellon University, "picoCTF," 2024. [Online]. Available: https://picoctf.org/

[87] DEF CON, "The DEF CON CTF," Aug. 2021. [Online]. Available: https://defcon.org/html/links/dc-ctf.html

[88] G. Vigna *et al.*, "Ten years of iCTF: The good, the bad, and the ugly," in *Proc. USENIX Summit Gaming, Games, Gamification Secur. Educ. (3GSE)*, San Diego, CA, USA, Aug. 2014.

[89] Shellphish, "The iCTF framework," Dec. 2021. [Online]. Available: https://github.com/shellphish/ictf-framework

[90] California Mathematics, Science, and Computer Science Partnership (CAL-MSCS), 2024. [Online]. Available: https://sites.google.com/sjcoe.net/cal-mscs/home/

[91] Cybersecurity and Infrastructure Security Agency, "Target rich, cyber poor: Strengthening our nation's critical infrastructure sectors." [Online]. Available: https://www.cisa.gov/news-events/news/target-rich-cyber-poor-strengthening-our-nations-critical-infrastructure-sectors (accessed Nov. 17, 2026).

[92] V. Willson, T. Monogue, and C. Malave, "First year comparative evaluation of the Texas A&M freshman integrated engineering program," in *Proc. Frontiers Educ. Conf. (FIE)*, 1995, pp. EAc13–EAc19.

[93] M. Hoit and M. Ohland, "The impact of a discipline-based introduction to engineering course on improving retention," *J. Eng. Educ.*, vol. 87, no. 1, pp. 79–85, 1998.

[94] D. W. Knight, L. E. Carlson, and J. F. Sullivan, "Staying in engineering: First-year projects course on student retention," in *Proc. ASEE Annu. Conf. Expo.*, 2003.

[95] Hart Research Associates, "It takes more than a major: Employer priorities for college learning and student success," Assoc. Amer. Colleges Univ., Washington, DC, USA, Tech. Rep., 2013.

[96] Hart Research Associates, "Falling short? College learning and career success," Assoc. Amer. Colleges Univ., Washington, DC, USA, Tech. Rep., 2015. [Online]. Available: https://www.aacu.org/leap/public-opinion-research/2015-survey-falling-short

[97] Chronicle of Higher Education, "The role of higher education in career development: Employer perceptions," 2012. [Online]. Available: http://www.chronicle.com/items/biz/pdf/EmployersSurvey.pdf

[98] National Academy of Sciences, National Academy of Engineering, and Institute of Medicine, *Enhancing the Postdoctoral Experience for Scientists and Engineers: A Guide for Postdoctoral Scholars, Advisers, Institutions, Funding Organizations, and Disciplinary Societies*. Washington, DC, USA: National Academies Press, 2000.

[99] National Academy of Sciences, National Academy of Engineering, and Institute of Medicine, *The Postdoctoral Experience Revisited*. Washington, DC, USA: National Academies Press, 2014.

[100] E. Trickel *et al.*, "Shell we play a game? CTF-as-a-service for security education," in *Proc. USENIX Workshop Adv. Secur. Educ. (ASE)*, Vancouver, BC, Canada, Aug. 2017.

[101] A. Doupé *et al.*, "Hit 'em where it hurts: A live security exercise on cyber situational awareness," in *Proc. Annu. Comput. Secur. Appl. Conf. (ACSAC)*, Orlando, FL, USA, Dec. 2011.

[102] O. A. Adedokun, A. B. Bessenbacher, L. C. Parker, L. L. Kirkham, and W. D. Burgess, "Research skills and STEM undergraduate research students' aspirations for research careers: Mediating effects of research self-efficacy," *J. Res. Sci. Teach.*, vol. 50, no. 8, pp. 940–951, 2013.

[103] L. J. Charleston, "A qualitative investigation of African Americans' decision to pursue computing science degrees: Implications for cultivating career choice and aspiration," *J. Diversity Higher Educ.*, vol. 5, no. 4, pp. 222–243, 2012.

[104] D. English and P. D. Umbach, "Graduate school choice: An examination of individual and institutional effects," *Rev. Higher Educ.*, vol. 39, no. 2, pp. 173–211, 2016.

[105] L. W. Perna, "Understanding the decision to enroll in graduate school: Sex and racial/ethnic group differences," *J. Higher Educ.*, vol. 75, no. 5, pp. 487–527, 2004.

[106] A. E. Leak, E. Sciaky, L. Lenaburg, J. A. Bianchini, and S. Scott, "Essential elements of collaboration: Understanding how chemistry graduate students experience collaboration through international research visits," *J. Chem. Educ.*, vol. 95, no. 5, pp. 749–757, 2018.

[107] V. Luna and L. Prieto, "Mentoring affirmations and interventions: A bridge to graduate school for Latina/o students," *J. Hispanic Higher Educ.*, vol. 8, no. 2, pp. 213–224, 2009.

[108] A. M. Wofford, L. J. Sax, K. L. George, D. Ramirez, and C. Nhien, "Advancing equity in graduate pathways: Examining the factors that sustain and develop computing graduate aspirations," *J. Higher Educ.*, vol. 93, no. 1, pp. 110–136, 2022.

[109] A. Alshahrani, I. Ross, and M. I. Wood, "Using social cognitive career theory to understand why students choose to study computer science," in *Proc. ACM Conf. Int. Comput. Educ. Res. (ICER)*, 2018, pp. 205–214.

[110] J. M. Blaney and A. M. Wofford, "Upward transfer student pathways in computing: Examining degree and career outcomes through structural equation modeling," *Community College J. Res. Pract.*, pp. 1–22, 2023.

[111] S. Cheryan, A. Master, and A. N. Meltzoff, "Cultural stereotypes as gatekeepers: Increasing girls' interest in computer science and engineering by diversifying stereotypes," *Front. Psychol.*, vol. 6, 2015.

[112] C. McCallum, "Giving back to the community: How African Americans envision utilizing their PhD," *J. Negro Educ.*, vol. 86, no. 2, pp. 138–153, 2017.

[113] A. M. Wofford, "Modeling the pathways to self-confidence for graduate school in computing," *Res. Higher Educ.*, vol. 62, no. 3, pp. 359–391, 2021.

[114] B. Qureshi, "Exploring the use of ChatGPT as a tool for learning and assessment in undergraduate computer science curriculum: Opportunities and challenges," in *Proc. 9th Int. Conf. E-Soc., e-Learn. e-Technol.*, 2023, pp. 7–13.

[115] N. M. Gürel, X. Qi, L. Rimanic, C. Zhang, and B. Li, "Knowledge enhanced machine learning pipeline against diverse adversarial attacks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 3976–3987.