Taking Bold **ACTION** to Bolster Cyberdefense

Cyberattacks have become increasingly common, sophisticated, and costly. Researchers at the new NSF-funded, UCSB-led ACTION Institute intend to team humans and AI to protect mission-critical systems and infrastructure.

ACTION figures: Institute director, Giovanni Vigna (right), and his co-PIs (from right), João Hespanha, Christopher Kruegel, and Ambuj Singh, lead an eleven-university collaboration to develop new AI-human partnerships intended to revolutionize cyberdefense.

The proposal that UC Santa Barbara researchers submitted to the National Science Foundation (NSF) for a grant to develop new ways of combating cyberattacks, with artificial intelligence (AI) as a main component, included a hypothetical attack scenario. In it, a group of individuals aligned with a hostile nation-state launch a sophisticated multiphase attack against key infrastructure elements of a fictional city: New Esperanza. The scenario is a chillingly realistic representation of how sophisticated hackers can gain access to inadequately defended cyberconnected systems.

The proposal succeeded, and last May, UCSB was named the lead institution in a five-year, $20 million NSF grant to pursue new approaches to cybersecurity linking humans to AI agents, and multiple agents to each other. UCSB computer science professor **Giovanni Vigna** is the institute's director. He is joined by fellow co-PIs (and UCSB professors) **Christopher Kruegel** (computer science), who has worked with Vigna on seminal research in the areas of intrusion detection, malware analysis, and threat intelligence; **Ambuj Singh** (computer science), a renowned expert on machine learning on networks and human-AI teaming; and **João Hespanha** (electrical and computer engineering), a world expert in control systems, game theory, and optimization. In addition, the NSF Institute for Agent-based Cyber Threat Intelligence and OperatioN (ACTION) brings together 21 other top AI researchers from ten other U.S. universities in a collaborative effort to develop revolutionary new forms of integrated cyberdefense.

Vigna describes the ACTION Institute researchers as "some of the very best people in AI and security, who have been at the forefront of expanding the foundations of AI, machine learning, game theory, and computer security." They and each of their institute colleagues will work primarily in one of eight highly integrated and interdependent research thrusts — four each in foundational AI and cybersecurity.

Paralyzing a City

The attackers in the New Esperanza scenario aim to create uncertainty and chaos by shutting down the city's water- and power-distribution infrastructure, which are controlled, respectively, by the Great Aqueduct and the Las Palomas power plant. The control systems for both are integrated with New Esperanza's smart-city system, which incorporates open-source software to monitor and distribute power, water, and other services.

The nation-state actors gather intelligence about the targets, identify open-source software used in the smart-city system, and then use false identities to contribute a vulnerable software component to the project, which goes undetected. They use credentials obtained from underground forums to connect to the virtual private network (VPN) of the aqueduct system, gain entry to various connected systems, introduce and exploit a vulnerability to obtain administrative access to the main server and upload a wiper malware component, all in ways beyond the ability of the systems to detect. After a few more steps, the attackers cause the power plant to cease operations,

> "There are simply not enough people to monitor what's happening in a network of mind-boggling complexity, make sense of it, and identify and resolve problems in a timely fashion."

such that the smart-city system cannot be controlled. Simultaneously, they activate malware that they installed, shutting down the aqueduct and blocking water flow to New Esperanza. The city is paralyzed, and chaos ensues.

Details of the attack included in the NSF proposal highlight multiple fail points at which suspicious or otherwise anomalous activity went undetected, exactly the kind of vulnerabilities that can bring down the operations of any connected entity that is inadequately protected. ACTION Institute researchers plan to bring forward innovations in AI and its application to cybersecurity that will protect critical infrastructure from sophisticated attacks like this one.

Fighting Back: Challenges of Time and Scale

Currently, the task of defending against cyberattacks depends largely on the skills, intuitions, and experience of human defenders, who must attend to all the elements of a typical cyberdefense life cycle: risk assessment and prevention, detection, attribution, and response and recovery.

As a result of the ever-increasing number, complexity, and sophistication of cyberthreats, however, the effectiveness of humans who staff the thousands of security operations centers (SOCs) at the nation's hospitals, financial institutions, government agencies, and other large connected entities can no longer respond with adequate speed or at sufficient scale to combat next-generation threats. There are simply not enough people, Vigna says, "to monitor what's happening in a network of mind-boggling complexity, make sense of it, and identify and resolve problems in a timely fashion.

HARDEN: UCSB Leads Project to Combat "Weird Machines"

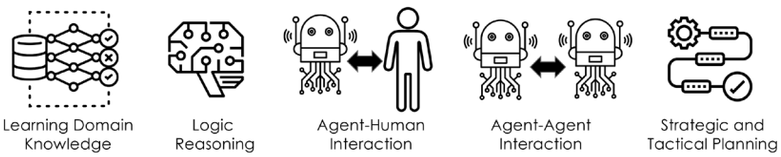
In another high-level cybersecurity project, UC Santa Barbara is one of two universities among eight groups (the other six are corporations) included in a new Defense Advanced Research Projects Agency (DARPA) project, called Hardening Development Toolchains Against Emergent Execution Engines (HARDEN). UCSB computer science (CS) professor **Tevfik Bultan** is the PI, with UCSB CS professors **Yu Feng, Christopher Kruegel,** and **Giovanni Vigna** as co-PIs. They are joined by collaborators at Purdue University. The four-year, $2.2 million project is intended to advance methods to improve defenses against a specific kind of attack at the firmware level.

Firmware, the lowest level of code in a computer, even beneath that of the operating system, executes critical functionalities and is susceptible to what are called *emergent behaviors*. Cyberattackers increasingly target firmware, which runs when computers boot up, in order to dodge security protections before they are activated. Compromising these basic building blocks of a computing system destroys the trustworthiness of a computer — or a device, such as a tablet used as an aircraft pilot's "electronic flight bag."

Emergent behaviors result in what are colloquially described as "weird machines," which means, essentially, that an attacker exploits flaws in a computer's code to compromise a feature and create unexpected behaviors, allowing the attacker to operate the system in ways never intended. They can then use that first compromised feature to attack and compromise another feature, and so on. This "compositional" method of accumulating compromised elements of the firmware — and the resulting emergent behaviors — can be hard to identify and is especially dangerous, because, first, it allows benign features built into the system by the manufacturer to be exploited by an attacker, and, second, while the emergent behaviors are ephemeral, they are robust, and the chains that drive them are portable between implementations created independently by different vendors.

"Emergent behaviors make a computer more susceptible to attack by allowing it to be used in a way it is not meant to be used," Bultan says. "We want to discover these kinds of attacks and mitigate them by hardening the system against them."

DARPA says that HARDEN aims to develop pioneering formal methods and automated software analysis to "deny hackers the ability to turn parts of modern computing systems against the whole."



Signposts to action: The ACTION Institute AI stack will allow intelligent agents to (from left) learn and reason about new facts, interact with humans and other AI agents, and engage in tactical and strategic planning in the face of uncertainty.

"Solving that time-and-scale problem will require automation," he adds, "but it has to be smart automation, and that means AI."

The ACTION Institute is part of a $140 million investment by the NSF, in collaboration with other federal agencies and stakeholders, to establish seven new National Artificial Intelligence Research Institutes, itself part of a broader federal effort to advance a cohesive national approach to AI-related opportunities and risks.

Says Vigna, "The ACTION Institute mission is to find new AI concepts and constructs that can be used to create new security applications that will change how mission-critical systems are protected against sophisticated, ever-changing security threats."

That will occur on two broad fronts: one is fundamental AI research — finding new ways for AI to model and reason about knowledge; the other is creating interaction and integration between and among humans and autonomous AI agents.

Stacking the Defense

ACTION Institute researchers aim to accomplish their mission by building a new AI stack, "a set of integrated tools that work together like a package that allows you to build AI-powered applications," Vigna explains. The AI stack will provide ways for intelligent agents to learn new facts and reason about them, communicate with humans and with each other, and support the planning of their actions.

These basic AI capabilities become the building blocks for developing security intelligent agents, such as agents that identify vulnerabilities in software before they are exploited, or intelligent agents that are able to suggest an effective remediation procedure after a breach has been detected.

One notable aspect of this AI stack is its focus on logical reasoning: While current AI approaches to cybersecurity mostly focus on machine learning (that is, the learning from large amounts of data), the vision brought forward by the ACTION Institute focuses on being able to apply deductive and inductive reasoning on what is observed in a computer

> "This new AI stack will need to operate in a world where attackers also use automation and AI to overcome cyberdefenses. Designing security systems must therefore involve reasoning about how the actions of one AI agent will affect the behavior of another agent."

network. This will support novel ways to understand the security posture of critical systems and deploy effective protections.

"This new AI stack will need to operate in a world where attackers will also use automation and AI to overcome cyberdefenses," João Hespanha explains. "Designing security systems must therefore involve reasoning about how the actions of one AI agent will affect the behavior of another. This type of reasoning is needed to make sure that whatever protection mechanisms we deploy to protect our systems do not create a completely new vulnerability."

Reasoning & Human-Agent Teams

Intelligent security agents, defined in the proposal as "[non-human] entities that employ reasoning, learning, and collaboration to perform one or more cybersecurity functions," will leverage the stack's capabilities to serve their functions in an uncertain, dynamic adversarial environment, with the agents following a new paradigm of continuous lifelong learning, both autonomously and in collaboration with human experts.

"We want the AI to continuously learn new facts, because computer networks are complex, evolving systems, and the intelligent agents need to continuously update their knowledge to be effective" Vigna says. "That capability is in its infancy right now, but work from the institute will bring it forward in an interesting way."

"Over time," the NSF proposal reads, "these intelligent security agents will become increasingly robust and effective as adversaries change modes of operation, more capable of composing defense strategies and tactical plans in the presence of uncertainty, more collaborative with each other and with humans for mutually complementary teaming, and better able to adapt to unfamiliar attacks."

The research is aimed at producing a major shift: "providing breakthroughs in AI necessary to evolve the current human-driven and human-paced security process into an agent-driven autonomous process… that continuously improves the security and resilience of computer systems… to ensure the confidentiality of sensitive data and the protection of critical services, saving billions of dollars and, in some cases, human lives."

"This concept of autonomous intelligent agents that are capable of reasoning, and, at the same time, focusing on security, is new," Vigna says. "Right now, there's nothing like that. There is no autonomous agent that is able to talk to other autonomous agents."

> "Responding to evolving threats requires reasoning and acting based on small amounts of data and adapting to untrainable and unspecified scenarios. In the case of security, humans and AI may have different perspectives on the implications of actions. This is where the synthesis of humans and AI becomes useful."

"AI agents are typically good at well-defined tasks when there is lots of training data," Ambuj Singh observes. "But responding to evolving threats requires reasoning and acting based on small amounts of data and adapting to untrainable and unspecified scenarios. In the case of security, humans and AI may have different perspectives on the implications of actions. This is where the synthesis of humans and AI becomes useful.

The basic idea behind the integrated approach, Singh adds, is that, "We need to have agents everywhere to prevent or repel an attack in time and at scale. We believe that the extensive domain knowledge, logic-based reasoning, human-agent, and agent-agent interactions enabled by our AI stack will provide all of those capabilities."

Trust, Ethics, and the AI Landscape

The work of developing the AI stack comes with tremendous challenges. For instance, Vigna says, "When you have agent-to-agent interaction, you have autonomous agents that are going around your network fixing things, and they have to talk to each other. If only one person programs all of them, it's easy, but an intelligent agent at UCSB might have to communicate with, say, an agent at UC Irvine or in a completely different realm, maybe at a financial

Humans working at security operations centers like this one, depicted via an AI-driven illustration app, can simply not keep up with the escalating scale of cyberattacks.

institution, about a concerning pattern of activity so that they can look for it.

"Before an agent can do that, however, we have to make sure that we preserve the privacy of the people involved by not disclosing, for example, that a specific human user went to a specific website. At the same time, we want to create some useful knowledge that can be used by other people can use to protect themselves. Properly configuring these agent-to-agent interactions to balance those needs is hugely important."

ACTION Institute researchers begin their work keenly aware of that challenge and an array of others associated with AI-enabled functionalities, from biases learned from existing datasets to "hallucinating" large-language models. Above all, Vigna says, "We want to have ethical AI. We don't want it to be making decisions that could cause harm — and not necessarily even physical harm; it could be something simple like having your computer cut out from the internet because an AI agent made the wrong decision. We want to be sure that decisions are made with a human in the loop, but in an efficient, targeted way that makes the best use of that person's capability."

"If you use AI wrong," Kruegel adds, "you can hurt entire classes of people, and that has occurred, so we have to be careful about what we encode in the agent's knowledge, what we learn from data, how we learn it, and how we align it to conform to the highest ethical values. Developing AI that is ethical and trustworthy is not an option; it's the only thing you can do, and it is ingrained in this community. Of course, AI is a tool, and a tool can be misused. That's why we have to be extremely careful."

With UCSB — home to the Center for Responsible Machine Learning — as the lead institution, ACTION Institute researchers will be focused on developing AI that is ethical and equitable every step of the way.

Collaboration and "Polarizing" Interest

The NSF established the seven new AI institutes simultaneously with the idea that the researchers in different domains would support each other and extend the value of their expertise through collaboration. "What the NSF wants, and what we also want is to deliver, in terms of research, more than what would result from giving twenty $1-million grants to twenty people," Vigna says. "We want something that comes out of the synergy of creating these cohorts of people from AI and from security and having them work together. The basic idea of our institute is to combine two cultures — one that is looking for new ways to do AI and another that is looking to use AI in new ways to improve security. We hope that by putting them in the same room, something amazing will result. Synergies will be really important to the success of this project."

Vigna hopes, too, that the institute will serve as a kind of North Star for AI-focused security research, providing a general direction in which to aim research done by people even beyond the institute who are involved in efforts that may be related, even if they are not entirely aligned.

"When you create an institute with a specific emphasis, you create almost a gravitational pull toward the topic that makes other people understand that this is important," Vigna explains. "I'm already seeing it. I might go to a conference about designing AI security, and people realize, 'Oh, so this is happening,' and they get pulled into it. I hope that the institute can become a nexus for both the AI and cybersecurity communities, polarizing interest and motivation around the topic, and aligning disparate interests."

A Stack at Market?

At the end of each year of the project, researchers will build increasingly sophisticated prototypes and use testing environments similar to that in the hypothetical New Esperanza model to test and demonstrate the stack's evolving capabilities. "Once you have a prototype that can demonstrate the abilities of what you've developed, it's much easier to transfer technology to industry, which is where it can have a real impact," Vigna notes. "Our ultimate goal is to demonstrate what can be accomplished by innovating both AI and cybersecurity; it's not our job to turn these ideas and prototypes into a product. We hope that the big security vendors will pick up the techniques and approaches we develop and transfer them to a commercial product. That would be a fantastic outcome."

Education, Workforce Development, Community Engagement

Mindful of the deepening presence of AI in every area of life, the resulting need to expand the pool of AI experts, and the fact, noted in the proposal, that "Early engagement is key to diversifying the STEM pipeline," ACTION Institute leaders have outlined innovative educational plans and workforce-development tools targeted at the K-12, undergraduate, graduate, and postgraduate levels.

The aim in terms of younger students, reads the NSF proposal, is to "nurture our youth's love for learning and cultivate their independent learning skills." The institute will also design and implement two yearly competitions centered around AI and security, one focused on high school students, and one devoted to undergraduate and graduate students.

These "Capture The Flag" competitions, which were pioneered by UCSB's Security Group and have been run by it for more than twenty years, have demonstrated their effectiveness in exciting students about the possibilities of combining cybersecurity and artificial intelligence. It is those students, many of whom are only in grade school now, who will play crucial roles in designing and implementing future versions of AI defenses "stacked" against sophisticated attacks.

Education is key to expanding the pipeline of AI-knowledgeable cybersecurity experts.