Creating a Scientific Foundation for Cyber Autonomy

Lujo Bauer, David Brumley, Joseph Calandrino, Nicolas Christin, Giulia Fanti,
Virgil Gligor, Bryan Parno, Jignesh Patel, Vyas Sekar, Justine Sherry
CyLab, Carnegie Mellon University

v1.0.0
March 20, 2026

Executive Summary

AI-driven cyber-offense capabilities are continuing to improve (e.g., Anthropic recently reported a cyber threat created and executed using AI tools) [9, 50] and system size and complexity continue to increase (e.g., with increasing deployments of agentic workflows in various aspects of enterprise workflows and AI-generated systems). In this setting, we argue that classical approaches to cybersecurity that rely on human timescales for exploration, detection, and mitigation will become increasingly inadequate. To defend our digital and cyberphysical infrastructures, we believe that cybersecurity defenses will have to be *autonomous* or partially autonomous and operate at *machine timescales*; we call this *Cyber Autonomy*, and make the case for developing the scientific foundations to realize this vision.

In particular, we argue it is necessary to rethink traditional security research and development, focusing on an *integrative system-wide approach* rather than siloed research explorations. By integrative, we mean not only looking at individual subcomponents, research challenges, and stakeholders, but also taking a holistic view on how the various components, stakeholders, and research outcomes will interact in a deployed operational environment. This entails: (1) creating capabilities for the design and implementation of end-to-end autonomous attack and defense systems; (2) devising frameworks for understanding interactions across multiple stakeholders including human operators, AI vendors, security vendors, and platform providers; and (3) developing foundations and empirical evaluations for rigorously understanding the interactions between diverse attack and defense systems and environments. By embracing such an integrative approach, we believe the community can create the foundations—algorithms, systems, datasets, and benchmarks—to advance and evaluate future AI-driven autonomous and semi-autonomous attack, defense, and operational capabilities. Admittedly, we have more questions than answers, and we likely have coverage gaps in the set of questions and even the types of questions we are asking. Our goal in writing this manifesto-style paper is a community-wide call for action to highlight research, development, and operational challenges that need to be overcome if we are to successfully meet the rising threats.

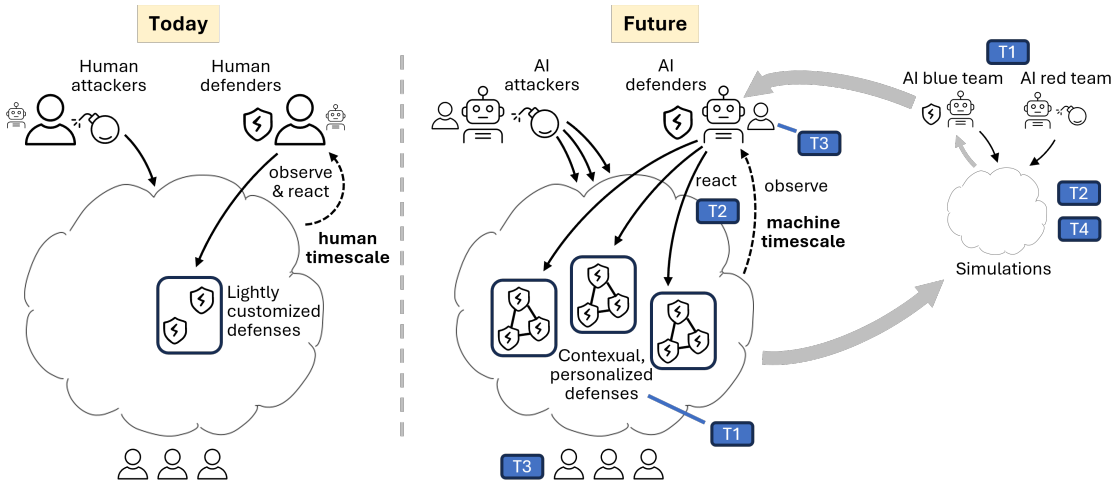

Figure 1: The contrast between today's human-timescale of responses to attacks and a future vision of much more autonomous cyber operations. The figure highlights a logical observe-orient-decide-act (OODA) loop that such largely autonomous workflows will employ and how it relates to the technical thrusts (T1: Algorithmic foundations; T2: Systems support; T3: Human factors; T4: Evaluation arena) we highlight in Figure 2.

1 Introduction

The world is at a pivotal moment in which advances in AI are dramatically changing the playing field in cybersecurity. Thanks in part to AI coding assistants, our ability to quickly build and deploy new features and systems far exceeds our ability to secure them. AI is concurrently changing how cyber threats evolve and has already created new capabilities for attackers. For instance, our early work at CyLab has shown the feasibility of autonomously generating exploits (e.g., [1]) and carrying out complex, multi-stage attacks (e.g., [77]), foreshadowing capabilities that are now appearing in the wild [8]. These new attacker capabilities, combined with the rate at which new code is being produced, mean that attacks will increase in number, velocity, intensity, and breadth. This will create an unprecedented level of risk for all types of organizations and networks, from non-profits and universities, to commercial enterprises, to government and critical infrastructure.

Setting: Imagine a hypothetical company ACME Enterprises. ACME Enterprises needs to secure their critical assets (e.g., proprietary data, critical machines) against existing threats and emerging autonomous attacks. A typical deployment of cybersecurity defenses will consist of capabilities to inform the OODA (observe-orient-decide-act) loop [66] in the Security Operations Center (SOC). The SOC's goal is to choose suitable security postures for various types of assets and scenarios based on the current operating context. To do so, the SOC requires the following capabilities: (1) *Observe:* Telemetry capabilities at hosts and network vantage points; (2) *Orient:* Data processing capabilities to make sense of the telemetry and provide useful contextual information to the SOC; (3) *Decide:* Offline and online emulation capabilities to run what-if analysis scenarios to inform the AI ML capabilities in the policy layer; (4) *Act:* Defense enforcement capabilities available at host and network vantage points and orchestrated effectively to implement the suggested actions to avoid, investigate, or remediate attacks.

Why now and why we need foundational research: For settings like our hypothetical ACME enterprise, we project that for the foreseeable future, system complexity will continue to grow and AI-powered attacks will lower the cost to launch and scale attacks against their infrastructures. The existing mechanisms, protocols, and processes used to secure enterprises are based on a "human attacker" mindset (see Figure 1). Today's security operations still significantly rely on manually predefined rules that grant or deny access based on simple heuristics and on defender input delivered on a human time scale. To tackle the emerging and growing threats, we will need several orders of magnitude (say 100×) improvements in our ability to defend systems at scale, at low cost, and against novel emerging threats. While speed, scale, and velocity are critical, we also need to tackle practical concerns with respect to trust, accountability, cost, vendor interactions, complex (and dynamic unbounded)[1] environments, among other things.

An integrative research methodology: To develop defenses capable of meeting such threats, we argue that we need a different paradigm for cybersecurity, development, and experimentation that departs from conventional siloed approaches. Today, plenty of excellent security research and development focuses on a specific piece of the puzzle, such as studying password usage (e.g., [58]), bug finding [29,31], or patching capabilities [54,55], in isolation, without holistically evaluating how the creation of a specific new capability would change the balance between attackers and defenders. While this hyperspecialized focus has been and will continue to be valuable, we argue that to create a rigorous foundation for Cyber Autonomy we need to "go back to the future" and take inspiration from the early days of computing and security. In particular, in the early days of computer security and computer systems, there was a more holistic approach and focus on building systems that were operational or at the very least functionally near operationalizable [22, 26, 51].

By integrative, we mean that developing the foundations for Cyber Autonomy includes, but needs to go beyond, discovering new zero-days (e.g., [34]), guardrails for prompt-injection attacks (e.g., [53]), or autonomously solving CTFs (e.g., [91]) and code challenges (e.g., [18, 92]). While research and development on these and many other building blocks is important and necessary, we argue that the research community needs to look at the problem holistically and create foundational abstractions, artifacts, datasets, and benchmarks, and evaluate them in realistic system-wide settings to move the proverbial needle. This *integrative philosophy* is characterized by three facets:

- *End-to-end designs and implementations of autonomous defense and red-team systems:*[2] Our example ACME SOC is in charge of defining an overall security posture of the entire system, not just of a subsystem or a piece of code. That is, they need mechanisms to defend the whole networked system against future autonomous attacks. As the security community knows, understanding the capabilities of future attackers is critical to ensuring a good defense. Thus, our ACME SOC also needs to invest in proactive end-to-end red-team systems to continuously stress-test their security postures.

- *Holistic implications for multiple stakeholders including human operators:* Our ACME SOC needs to manage the entire networked system, potentially using both in-house capabilities and

[1] For instance, many enterprises cannot have a well defined perimeter or do not have mechanisms to completely enumerate assets.

[2] We note that this is different from the end-to-end arguments in system design that argue for placing functionality in the end-host stack rather than intermediate network devices, unless there is a significant performance advantage to doing so [74]. Rather, we use the term end-to-end in the sense of designing, implementing, and evaluating operationalizable or near-operationalizable solutions that consider the system, workloads, users, and operators holistically [39].

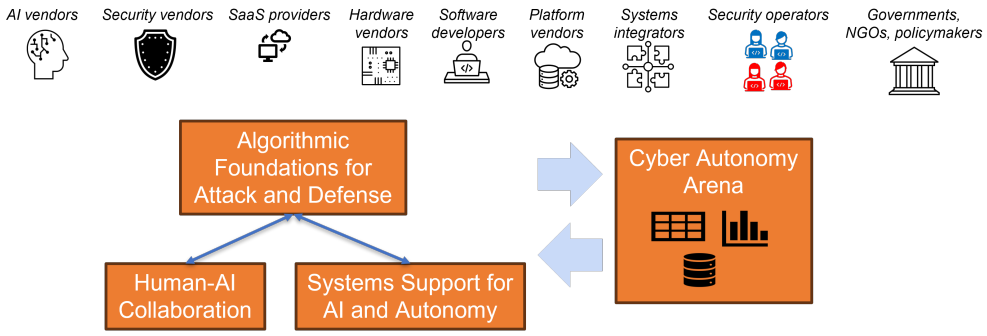

Figure 2: High-level vision of the key technical thrusts in Cyber Autonomy and the stakeholders who will benefit from the advances.

solutions from multiple vendors (e.g., AI/ML vendors, SIEM vendors, network vendors, endpoint security vendors) to achieve their goals. Solutions for Cyber Autonomy should consider the practical ecosystem realities and inform the interfaces, collaborations, and interactions across multiple stakeholders. Finally, even with autonomous operations, we envision human operators will still need to be accountable for agents acting on their behalf and thus operators will need corresponding tools to explain, inspect, and verify AI-cyber capabilities.

- *Rigorously understanding the interactions between diverse attack and defense systems in realistic environments:* To inform their deployment and security postures, the ACME SOC operators will need to make informed judgments on how different attack and defense solutions perform in *their specific environments*. Rather than rely on guesswork or on public benchmarks that may not reflect the specific environments, the SOC operators will need capabilities to evaluate offered vendor components and integrated systems rigorously in their own setting. More generally, the overall ecosystem of stakeholders will need evaluation capabilities analogous to AI evaluations and benchmarks (e.g., [43, 48, 79]). Our ACME SOC is in charge of delivering an overall security posture for the entire system, not just a subsystem. That is, they need mechanisms to defend the whole networked system against future autonomous attacks. As the security community knows, understanding the capabilities of future attackers is critical to ensure good defense. Thus, our ACME SOC also needs to invest in proactive red-team systems to continuously stress test their security postures.

This integrative philosophy permeates the research thrusts areas that we outline next.

Technical thrusts: In this whitepaper, we identify the need for interdisciplinary advances along four interconnected technical thrusts as shown in Figure 2. While we logically separate these areas for ease of exposition, they are fundamentally interconnected in line with our integrative systems philosophy. For instance, to be useful, the algorithmic foundations (Thrust 1) rely on telemetry, data processing, and orchestration advances from the systems foundations (Thrust 2). Similarly, both the algorithmic and systems foundations will benefit from the workloads and datasets created as part of the evaluation arena (Thrust 4). In a similar vein, the human-AI collaborative capabilities (Thrust 3) will aid and inform the algorithmic, systems, and arena advances. These advances will yield cross-cutting benefits across a wide range of stakeholders: developers, AI vendors, security

vendors, SaaS providers, hardware providers, platform vendors, system integrators, as well as policy makers and policy advisors.

- *Thrust 1: Advance the algorithmic foundations for end-to-end autonomous defense and red-teaming capabilities*

  As AI-driven attacks become faster and more effective, we need defenses that react faster, more effectively, and sometimes, but not always, with less human oversight [8]. To help defenses stay ahead of AI-driven threats, we argue that we need to accelerate the development of autonomous and semi-autonomous defense techniques, including white-hat red-team and cyber-offense capabilities to enable proactive defenses. This acceleration can also be helped by lowering the barrier for experimentation with and design of such algorithms, enabling a wider community of builders (and AI systems) to contribute novel ideas [19].

- *Thrust 2: Develop new data-plane and control-plane systems in support of cyber autonomy*

  To power future autonomous cyber defenses, we will need novel systems foundations (e.g., real-time data processing, novel AI-driven log processing capabilities, and software-defined data- and control-plane capabilities) [30, 32, 89, 90]. The key challenge here is one of scale, cost, and responsiveness to keep up with the machine-timescales of attacks and defenses.

- *Thrust 3: Understand and accommodate human-factors constraints in designing, implementing, and operationalizing autonomous defenses*

  While we need to advance the algorithmic and systems foundations, we also envision the need to advance our understanding of the human-AI interactions across multiple stakeholders. This includes creating a foundational understanding of the role of human operators in autonomous defenses, and building in explainability and guardrails from the get-go. In addition, we also need to create foundations for upskilling a future cyber workforce and creative mechanisms for sharing insights, datasets, and best practices (e.g., analogous to Indicator-of-Compromise sharing in sector-specific ISACs [2]).

- *Thrust 4: Create and maintain a neutral and broadly available evaluation platform or "Cyber Autonomy Arena" for empirical grounding*

  To have a common empirical grounding to advance our understanding and evaluate emerging capabilities, we argue that we need to design and maintain a community-wide *Cyber Autonomy Arena* to enable objective and realistic evaluations of novel attack and defense capabilities. In addition, this arena or vendor-neutral playground will serve as a platform to create datasets and best practices that enable continuous improvements for the artifacts in the previous thrusts.

Cross-cutting challenges: As we formulate these technical thrusts, we note several cross-cutting practical challenges across all of these thrusts. A first concern is *cost and sustainability*; are future autonomous solutions economically viable and sustainable. A second concern is *vendor lockin* as the gap between locked and open AI models widens; i.e., is it possible to achieve the benefits with open and open-source solutions or in vendor-agnostic ways. A third concern relates to *practical enterprise realities and loci of control*; many enterprises may not even know what assets they have or not have control over devices that interact with their assets. A final and overarching concern

is *balancing human cognition and automation*; while we envision many parts of the operational workflows can get automated, mission critical decisions will still need human-in-the-loop decision making.

Risks: A potential risk stemming from a research agenda that proactively considers novel threats is that evildoers can also learn from potential developments. We argue that supporting open, innovative, and foundational research is the only way to stay ahead of the curve; "security by obscurity" was never a safe option and will not be one in the AI era. A second risk is that academia trails the state of practice. To this end, we will need to actively engage key industry partners to make sure that the research is focused on medium- and long-term foundational problems, while at the same time being driven by near-term feedback from practitioners, including from interactions that include producing open-source artifacts and data sets in the short term.

2 Research Thrusts and Challenges

We frame four broad research thrusts and identify subareas, or tasks, that we believe need further exploration; we also pose some open questions for the research community. We say this with one very important caveat—we intend this articulation of questions primarily as a means to spark a broader discussion and debate. We do not expect that there will be universal agreement that these are the right high-priority topics nor do we intend to claim that the set we propose is comprehensive.

2.1 Algorithmic Foundations for Autonomous Defensive and Red-teaming Capabilities

---

Thrust 1: ***Advance the algorithmic foundations for end-to-end autonomous defense and red-teaming capabilities.***

In early work at CyLab, we showed that AI-driven red teams equipped with new abstractions can autonomously execute complex multi-stage attacks against realistic networks within minutes [77]. Similarly, early work on defensive tactics showed that deceptive strategies, if deployed correctly, can slow down and help defeat some of these AI-driven attackers [78].

We believe that such efforts only scratch the surface of what can and will be possible with future AI and security advances. Cyber defenses that exhibit dynamic behavior typically do so as a result of selecting from among statically pre-defined options or using off-the-shelf AI tools (like classifiers or LLMs) to guide attacks or defenses along previously observed paths. We believe that future attacks will be more dynamic and unpredictable (at human scale), and so will require similarly dynamic, flexible, and automated defenses. Hence, we believe that the cybersecurity community's tradition of discovering and studying incipient attacks [14] will be more important than ever, except that it is now novel *AI-enabled attack techniques* (i.e., individual attack or defense moves) and novel *AI-enabled, adaptive attack strategies* that we need to foresee, understand, and build defenses against.

Our overarching goal in this thrust is to understand and improve AI and agentic capabilities for offense and defense, including understanding cases in which defenses have relied on human time and effort in ways that are now unacceptably slow and costly in the face of emerging AI capabilities (a point we also explore in Thrust 3). The thrusts are interrelated, and many of the goals of this one

cannot be accomplished without advances in systems capabilities (Thrust 2). In addition, many specific concerns are shared across multiple thrusts, such as the need for resource efficiency and avoiding vendor lock in.

Task 1.1: *Investigate the limits of existing AI capabilities (including LLMs) as a basis for autonomous defense and red-teaming capabilities.*

AI has been used for many cyber defense tools, from spam and intrusion detection [26] to malware detection [76], finding vulnerabilities in source code [1, 18], and penetration testing [5, 24, 45]. These uses of AI typically have one or more properties that make it uncertain how effective they would be against AI-powered dynamic, machine-timescale attacks. For instance, ML models may require periodic retraining and may be successful only when tested against previously seen attacks.

This characteristic is particularly relevant for security solutions that leverage LLMs, e.g., for tasks like bug-finding or red teaming (e.g., [83]). The LLM-powered tools are often effective at achieving their goals, but it is frequently unclear to what extent this success is due to the LLMs having been trained on the exact instances of the problems they are asked to solve. And because frontier LLMs' capabilities often increase substantially (though not monotonically) from generation to generation, the effectiveness of many tools is often evaluated only with respect to specific versions of specific vendors' LLMs, leaving uncertainty about the tools' utility with other vendors' LLMs or future versions of LLMs.

Hence, we believe the first set of tasks related to building better AI-powered defenses is to develop a more foundational understanding of the limits of current tools. How often do current state-of-the-art defensive tools achieve excellent performance only in contexts that their constituent models have encountered in training? How do they fare against truly novel attacks and attacks that adapt at machine timescales? How fundamentally novel can new attacks be compared to the attacks that are used as part of the training data of existing tools? For tools built on LLMs and other large models, what specific properties of the models or training data are responsible for the tools' success? What approaches to leveraging AI components generalize better across different vendors' offerings? What are the cost-benefit tradeoffs of using LLM-based tools, e.g., taking into account their inference latency, resource requirements, and potential additional third-party dependencies? To what extent can the apparent benefits of using large, vendor-specific LLMs be achieved by more lightweight, open-source models?

Task 1.2: *Create new AI-enabled strategy capabilities and attendant abstractions for defense and red-teaming.*

To create the next generation of more flexible AI-based defenses and red-teaming tools, we believe advances are needed both in strategy (how tools marshal or assemble a set of predefined steps or moves into an effective attack or defense tailored to a specific adversary) and in techniques (the atomic steps that are the building blocks of an attack or defense strategy; we use "techniques" roughly consistently with the MITRE ATT&CK framework's usage of the term) [23].

In regards to improving tools' strategic capabilities, we believe that the next step in their evolution is to develop defenses and red-team tools that do fundamentally more than select from among options pre-defined by experts. Reinforcement learning (RL) has been underexplored as a method to allow security tools to develop strategies that go beyond what human experts have

envisioned or that could automatically incorporate the behavior and impacts of attacks that had not yet been imagined when the tools were created. We believe that RL and other AI techniques that have not been consistently explored in cybersecurity will be among the key drivers of advances in attack and defense tools' strategic capabilities. As one example, we envision the need to extend existing ideas from game theoretic modeling in cybersecurity to be applied in a truly integrative fashion to end-to-end realistic scenarios (e.g., [13, 57]), and empirically validate the theoretical predictions and findings (Thrust 4).

We expect that many future defenses will consist of autonomous agents that will interact among themselves, not just with human operators. Hence, a substantial concern arises in the composition of autonomous agent actions; e.g., provable establishment of shared responsibility, of conflict resolution, and accountability of autonomous agents. Questions regarding agent trustworthiness [72], not merely resiliency to a growing number of attacks in vulnerable systems and networks, should be answered before deployment and hence agent accountability can be established. Large classes of security problems arise from agent interactions, e.g., the security of emergent properties in ad-hoc networks [36]. Some such properties are decidable and can be established computationally, e.g., secure routing and connectivity, detection of node-replication attacks [67]. Overall, effectively securing end-to-end systems will require securing these autonomous agent interactions.

Exploring these new AI-enabled strategy capabilities will also require solving challenges in gathering data quickly and at scale (Thrust 2), devising methods and incentives to share information among stakeholders, and understanding how to incorporate human input and oversight into automated tools (Thrust 3).

Task 1.3: *Create new AI-enabled techniques (i.e., individual actions) for risk assessment and proactive risk mitigation.*

Whereas the previous task focused on better using existing techniques (i.e., better strategies built from known actions), a separate task is to improve those techniques and to devise new ones to achieve similar goals. In recent years, AI-enabled tools have demonstrated promise in creating new instances of specific techniques, such as finding 0-day vulnerabilities [15, 31] and creating exploits for such vulnerabilies [1], and even automatically creating patches [54, 83]. While some such tools have already become a useful part of broader defense or red-teaming toolkits [18], we believe that there yet remains significant untapped potential to create much more capable actions in many areas, including for code analysis and repair tools that only minimally require humans; for reconnaissance; for deceptive defenses; for remediation actions (e.g., quarantine).

As AI-enabled capabilities change how systems are built and defended, it will be particularly important to balance innovation with a careful assessment of potential unintended side-effects. For example, AI code assistants have been shown to significantly increase the ability of programmers to complete many tasks, including related to security; but recent work has also shown that this performance may come at the expense of creating new bugs, both obvious and subtle [28, 47, 68, 95], security vulnerabilities [69], and increased technical debt [42]. All of these side effects, if not contained, could lead to the creation of systems that are fragile in new and unexpected ways.

One way to enable innovation while avoiding unintended side-effects is to add guardrails to new techniques from the outset, so that AI-driven behaviors are constrained. How to do this, and how to do it while allowing sufficient freedom to generate *beneficial* unforeseen solutions is one of many open research problems. In the context of code generation, however, we see promise in techniques from the formal methods community, which would allow humans to provide specifications for the

code they want, and then require the AI to produce both the code *and* a proof that the code meets the specification [87]. Historically, producing such proofs required considerable human effort, but LLMs are increasingly capable of doing so autonomously [88]. Research and investment in this direction could lead to AI-generated code that exceeds the quality of traditional human-generated code.

2.2 Develop new data-plane and control-plane capabilities in support of cyber autonomy

Thrust 2: ***Advance the state of art of computing systems for data processing, network management, and control as we move toward a realm of autonomous operations.***

We have made significant advances in software-defined orchestration [32], telemetry collection and processing [80, 90], and experimentation capabilities [3, 4]. Autonomous, machine-timescale defense systems will need corresponding advances in infrastructure capabilities and novel telemetry collection and processing, computing, and communication systems advances to enable the autonomous capabilities that arise out of the previous research thrust.

This will naturally push the boundaries of existing capabilities on several dimensions: (1) scalable telemetry and log analytics to achieve low cost, high scale, and low latency [10, 90]; (2) novel software-defined data-plane and control-plane capabilities for real-time policy enforcement across complex environments [32]; (3) novel sandbox, staging, and verification systems for the security teams to be able to quantify and verify the risks involved with deployment [86]; and (4) novel capabilities for advancing emulation systems for what-if analysis and "sim2real" (e.g., [44]) datasets for training models. We elaborate on each of these challenges next.

Task 2.1: *Design reliable and scalable real-time telemetry and data analytics capabilities that can take in real-time telemetry and other enterprise contextual data to inform autonomous defenses.*

Even today, security systems generate enormous amounts of traces and log data. Finding interesting patterns is a needle-in-the-haystack problem that entails significant cost, complexity, and blind spots. Defenders face an overwhelming volume of alerts, with extremely low signal-to-noise ratios and average response times measured in weeks. Existing approaches for telemetry collection and processing do not scale to the complexity or automation level of next-generation threats.

While there are many concurrent efforts to build new advanced instrumentation capabilities (e.g., [46]), AI-driven log processing (e.g., [94]), telemetry analytics (e.g., [93]), and AI-driven SOC triage playbooks (e.g., [12]), there are still fundamental gaps in creating *expressive and efficient* solutions that can simultaneously achieve *high accuracy, low cost, and low latency at high scale.* Similarly, we need mechanisms to analyze large volumes of historical data to continuously understand failure patterns to improve capabilities in a self-learning loop (e.g., [17]).

At a high level, existing frameworks for data collection, storage, and processing have largely been derived from a classical relational model of data processing [16,90] intended for human consumption. However, the data platforms behind these frameworks are increasingly queried and consumed by autonomous agents that iteratively discover and interpret underlying data, rather than by humans who manually inspect results and refine queries [52]. We need to understand to what extent existing

data management methods can support the expressivity and efficiency needs of autonomous and agentic security operations in the future and develop new foundations for such future workloads (e.g., [30, 41, 60]).

We envision the need for novel abstractions and systems for telemetry collection, transmission, storage, and analytics processing that can simultaneously achieve high accuracy, low cost, high scale, and low latency. This raises a number of challenging questions. For instance, can existing data warehouse systems handle the stateful and context-aware monitoring needs of autonomous security systems? What kinds of expressive and efficient data analysis foundations do we need to handle future security-specific query and analysis workloads to balance cost, scale, and efficiency, without sacrificing accuracy? What intermediate representations do we need to transform raw telemetry into structured context and state that both analysts and agents can reason about, and connect this with relevant enterprise context (assets, identity, vulnerabilities, and historical behaviors)?

Task 2.2: *Design and evaluate advanced and contextual software-defined control plane orchestration systems to support autonomous operations.*

Today, many defensive capabilities rely on relatively static and coarse rules (e.g., static firewall policies or quarantining compromised hosts). As defenses become more autonomous, we see both novel opportunities and novel challenges. Unlike human operators, autonomous defenses could be better equipped to deploy more fine-grained and dynamic orchestration (e.g., triage, analysis, response, recovery) and may use more advanced capabilities than the coarse ones today.

For example, we can imagine more personalized, stateful, and context-aware guardians for every single host or application orchestrated by a central SOC [32, 89]. Similarly, while deception today is used in a fairly narrow use case (e.g., honeypots for intel), we can imagine autonomous defenses dynamically deploying novel deception capabilities to deter, slow down, and distract AI attackers [6, 33, 85].

Supporting such an architecture for real-time explainable responses and global policy enforcement across complex environments raises new questions and will need new advances in control and data plane capabilities. For instance, what dynamic host- and network-based monitoring and enforcement capabilities are needed to enable advanced autonomous and personalized defenses? Can existing software-defined orchestration controllers handle the real-time responsiveness needs of future defenses? How should the capabilities at the host and network layer be exposed to AI planning agents? How can we do this in an efficient and scalable way for large deployments?

Task 2.3: *Develop systems with built-in guardrails and protections from autonomous systems misfiring.*

Analogously to outsourced management, using autonomous systems for red-teaming, patching, remediation, and system management may incur new risks due to (un)intentional policy violations introduced by agents. When rolling out AI defender agents, it is important for the security team to be able to quantify and verify the risks involved with deployment. This raises interesting challenges for policy frameworks, and testing and verification systems, as well as system design. For instance: can existing testing and verification mechanisms test for potentially non-determinstic behaviors in such complex autonomous operations in mission critical settings [75]? Should we design "verifiable-by-construction" abstractions and systems for expressing defense and red-teaming approaches [86]?

2.3 Human Factors and Stakeholders

Thrust 3: ***Understand and accommodate human-factors constraints—and take advantage of operator expertise—in designing, implementing, and operationalizing autonomous defenses.***

Although some future AI-powered defenses may be completely autonomous, we envision that most benefit will come from tools that keep humans in the loop, whether to gain human authorization or seek human insight at critical points, or because the tools' purpose is to magnify human effort rather than replace it.

Task 3.1: *Understand and accommodate operator requirements and needs for deploying autonomous and near-autonomous red teams and defenses.*

A common reaction by practitioners to the suggestion that red-team exercises against real enterprise networks should be entirely automated is to push back, asking for automated tools to allow operators to examine proposed actions and to confirm that they won't harm the network that is being tested. We envision that most AI-enabled defenses or defense tools, even when semi-autonomous, will need to occasionally or regularly interact with human operators. In addition to interactions that are about seeking permission, we expect that operators will also often wish to give instructions that can help better direct the efforts of semi-autonomous tools, e.g., to focus a red-team exercise on a particular sequence of steps that will test newly deployed defenses.

Yet other interactions between AI agents and operators might serve to overcome the limitations of AI agents, which will lack the contextual awareness (i.e., conscious cognition) to understand the non-computational effects of attacks or defenses at machine speed, in real time. For example, emergence of some security properties cannot be decided computationally; e.g., reaching a future network state from an arbitrary state with a given set of commands [37, 40], establishing trust in networks of humans and computers [38]. Hence, an overarching challenge is to design cyber security mechanisms that augment autonomous AI agents with human operator decisions to achieve effective defenses in enterprise networks.

Substantial additional research is needed to determine when human operators should be consulted, what information needs to be presented to them to make decisions, and how that information should be structured. The answers to these questions will often be parameterized by human cognition and knowledge, but the needed interactions will sometimes be required by regulations or business processes. A critical part of providing sufficient information to operators will be *explaining* specific AI-driven actions and decisions. This has already been a topic of a large body of work (see [27]), including in the cybersecurity domain (see [62]). However, many AI algorithms remain resistant to useful explanations for their decisions being derived. More research is needed to improve these capabilities; alternately, choices of AI building blocks might need to be made on the basis of the ability of those building blocks to support adequate explanations.

Understanding how semi-autonomous tools should interface with operators goes beyond individual interactions: tools will need to instill in their operators a comprehensive understanding of what they have accomplished thus far during an engagement and how the specific touch-points that involve operators are necessary for success. These topics, sometimes discussed under the moniker *human-machine teaming*, have already been examined in prior research. However, understanding

them well enough for that understanding to be reflected in effective tool features is still an open question.

Task 3.2: *Develop tools that improve the effectiveness of cyber defense operators and researchers.*

Even in a future where autonomous defense tools are necessary and ubiquitous, many defense measures will still be designed and implemented by people. A challenge in developing new defenses is that they can be complex and tedious to implement, e.g., they may require new telemetry infrastructure and may need to be built up from low-level APIs [73]. Recent work has shown, for example, that even simple new deception defenses can take thousands of lines of code to program and that adapting their implementations to even slightly different environments is laborious and error prone [78]. We believe that new programming and operational tools are needed to make it easier for defenders and researchers to quickly design and implement new defensive strategies.

Such future tools may enhance security experts' design and development capabilities by offering them more convenient abstractions specifically geared to ease the creation of security mechanisms, such as our preliminary work has already suggested for the design of modular, easily extensible deception techniques [78]. It may also involve the development of vibe coding agents and harness engineering techniques [56] that are specific to programming and deploying cyber defenses.

Task 3.3: *Devise frameworks and incentives for sharing environments, datasets, attack scenarios, and best practices.*

Both ML- and even non-ML-based security tools need often high volumes of data for training and configuration [21, 59]. Without adequate data or fine-tuned models or configurations, potentially useful defense tools might be ineffective. Even prior to the rise of AI-enabled defenses, this has been a problem: For example, password strength estimators need training data or expertly designed configuration files, neither of which are included with the tools as distributed; instead, these are frequently considered the "special sauce" of individual experts.

As AI-based defenses become increasingly important, widely sharing the data, trained models, tool configurations, and tool evaluation results becomes imperative for defenses to be successful. This is particularly the case because individual organizations seeking to defend themselves might not themselves observe in detail many instances of the harmful behaviors they seek to protect against. This problem has also been recognized in other industries, leading to, e.g., the Department of Transportation to require car manufacturers and operators to share incident data from cars that use automated driving systems or driving aids [63].

Ensuring that cyber-defense-related artifacts can be shared freely enough to enable robust defenses is a socio-technical problem. Additional research can help create technical means to share artifacts in an integrity- and privacy-preserving fashion. Can we build models that avoid memorizing their training data and so can be shared without fear that they will divulge business secrets? Can we create methods to share tool test results so that they cannot be falsified? Can we design tools so that their configurations or components of their configurations are sufficiently separate from the details of organizational networks that sharing the configurations won't leak business secrets?

Beyond the technical innovation, we believe that creating regulatory or other incentive-based structures to promote sharing will be necessary. Not all such structures need to be imposed via external requirements; we may be able to design cooperative defense mechanisms that explicitly

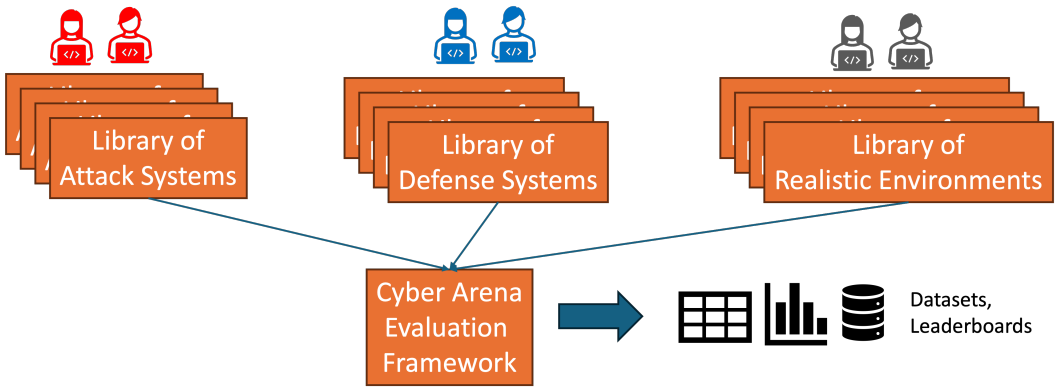

Figure 3: The final thrust is creating and maintaining a neutral Cyber Autonomy Arena enabling the broader community to contribute and evaluate novel ideas, and advance the science via datasets and benchmarks.

incentivize enterprises to join because doing so improves their own defensive posture.

2.4 Cyber Autonomy Arena for Capability Evaluation and Advancement

---

Thrust 4: ***Create and maintain a neutral, end-to-end arena for evaluating emerging AI offense and defense capabilities in realistic environments.***

Many disciplines in computing have been accelerated by the creation of community benchmarks, leaderboards, and tournaments (e.g., [11, 25, 43, 61, 71]). For example, in computer vision, ImageNet [25] dramatically accelerated the rigor of algorithm evaluation. In the systems community, the creation of benchmarks such as SPEC [43] and TPC [71] provided a common ground for scientific advances. In the AI domain, the emergence of benchmarks has provided a basis for community-wide evaluations (e.g., [20, 48, 70]).

As a security community, we need to have a grounded empirical understanding of the capabilities of attackers, defenders, network systems and vendor tools. We argue that there is a critical need for a leaderboard for evaluating *autonomous* agentic AI systems in network security.

Recall that a central tenet of our approach is the integrative philosophy. With respect to empirical understanding, the manifests on two dimensions. First, we need to evaluate end-to-end *attack systems vs. defense systems* rather than attack or defense tools in isolation. Second, we believe that understanding the *competitive interplay* between attack and defense systems in *realistic* environments is key to advancing the science of security. Unfortunately, existing efforts have key limitations, e.g., they focus on CTF-style challenges or datasets for evaluating individual attack and defense components (e.g., [76, 82]).

Today, cyber ranges for network security evaluations are limited in scale, diversity, and extensibility [33, 49, 64, 81]. We need abstractions and systems that enable creating diverse and realistic network security scenarios. We will need to address key challenges in scenario realism and diversity, and also mechanisms to accelerate the design of new challenge environments as well as novel

strategies for attack and defense. We envision several research and operational challenges in developing and maintaining such a system. For instance, can we create an open and extensible set of benchmark challenges and dynamically update them to avoid memorization where benchmarks get saturated over time [79, 96]? How can we customize these evaluation settings for diverse domain specific needs; e.g., industrial control systems, enterprise settings, cloud microservices? Can we incentivize and collaborate with vertical-specific ISACs to create reliable benchmarks? Given some environments and situations may be sensitive and vendors may not reveal their proprietary ideas, can we create zero-knowledge sharing mechanisms (e.g., [35]) for sharing results from private environments that can nevertheless inform the public benchmarks and overall policy?

3 Call For Action

Our vision here is that this research endeavor and community effort will create foundational algorithms, systems, open source artifacts, and datasets for advancing the theory and practice of Cyber Autonomy. We end with a call for action for various stakeholders and how they can contribute:

- *Academics:* While continuing to focus on advancing the science of individual subsystems, we would urge academics to address the integrative posture more seriously and also come together better as a community to create shared open artifacts and benchmarks rather than working in silos.
- *AI-model vendors:* AI-model vendors have already started doing cyber evaluations of the model capabilities [7, 65], sometimes with open benchmarks but also with private ones. We urge model vendors to work with academia and other stakeholders to make their products available for non-commercial evaluation in end-to-end settings.
- *Security-solution vendors:* Security vendors can contribute to the public benchmarks and realistic evaluations, and also provide neutral access to closed-source solutions for end-to-end evaluations to academics and non-commercial stakeholders.
- *Infrastructure vendors:* Running such community benchmarks is a non-trivial task both in terms of the capital and operating costs. Providing access to infrastructure (e.g., cloud, GPU hours) or support to keep the cyber autonomy arena operational will be extremely valuable. Infrastructure vendors are also likely a constant target of novel attacks so they can also provide valuable intelligence on emerging threats both against the infrastructure and their customers' infrastructures.
- *Federal agencies and philanthropic foundations:* The algorithmic, system, human factors, and data advances should be viewed as a digital public good [84] and there should be more community-wide investment in creating and advancing the foundations described here.
- *Policy think tanks:* While many of the advances will remain technical, policy think tanks serve as a critical bridge between the technical advances and the socioeconomic impacts and help articulate the implications of Cyber Autonomy to relevant decision makers.
- *Security operators:* There is a significant disconnect between the state of research and state of practice when it comes to operational science and wisdom. Operators can help better

inform the research by documenting best practices and novel operational strategies. Analogous to sharing of IoCs, sharing playbooks in reacting to incidents and hosting students and researchers in SOCs as trainees and partners will help better bridge the theory-practice gap.

- *Software developers:* AI is already changing the software development life cycle, but as we move toward autonomous agentic workflows in the SecOps lifecycle as well, it is also worthwhile for developers to think about the end consumers of their artifacts; e.g., making tools more amenable to agentic systems for test, verification, and usage and also potentially thinking about more declarative abstractions to help the human actors make sense of intents.

References

[1] Mayhem Security. https://www.mayhem.security/.

[2] National Council of ISACs. https://www.nationalisacs.org/.

[3] Project Vera. https://github.com/project-vera/vera.

[4] A. Agrawal, N. Kedia, J. Mohan, A. Panwar, N. Kwatra, B. Gulavani, R. Ramjee, and A. Tumanov. Vidur: A large-scale simulation framework for LLM inference. *arXiv preprint arXiv:2405.05465*, 2024.

[5] H. S. Al-Sinani and C. J. Mitchell. PenTest++: Elevating ethical hacking with AI and automation. *arXiv:2502.09484*, 2025.

[6] M. H. Almeshekah and E. H. Spafford. Planning and integrating deception into computer security defenses. In *Proceedings of the 2014 New Security Paradigms Workshop*, pages 127–138, 2014.

[7] Anthropic. Anthropic Claude 4 System Card. https://www.anthropic.com/claude-4-system-card, 2025.

[8] Anthropic. Disrupting the first reported AI-orchestrated cyber espionage campaign. https://www.anthropic.com/news/disrupting-AI-espionage, Nov. 2025.

[9] A. Anurin, J. Ng, K. Schaffer, J. Schreiber, and E. Kran. Catastrophic cyber capabilities benchmark (3CB): Robustly evaluating LLM agent cyber offense capabilities. *arXiv:2410.09114*, 2024.

[10] M. Armbrust, R. S. Xin, C. Lian, Y. Huai, D. Liu, J. K. Bradley, X. Meng, T. Kaftan, M. J. Franklin, A. Ghodsi, and M. Zaharia. Spark SQL: Relational data processing in Spark. In *Proceedings of the ACM SIGMOD*, 2015.

[11] R. Axelrod and W. D. Hamilton. The evolution of cooperation. *Science*, 211(4489):1390–1396, 1981.

[12] K. Banstola, F. A. Faisal, and X. Ou. Experiences of using agentic AI to fill tooling gaps in a security operations center. In *Proceedings of the Workshop on Security Operations Centers (WoSOC) at the Network and Distributed System Security Symposium (NDSS)*, 2026.

[13] T. Bao, Y. Shoshitaishvili, R. Wang, C. Kruegel, G. Vigna, and D. Brumley. How shall we play a game?: A game-theoretical model for cyber-warfare games. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 7–21, 2017.

[14] D. Basin and S. Capkun. The research value of publishing attacks. *Commun. ACM*, 55(11):22–24, Nov. 2012.

[15] L. Bilge and T. Dumitraş. Before we knew it: An empirical study of zero-day attacks in the real world. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*, 2012.

[16] P. Carbone, A. Katsifodimos, S. Ewen, V. Markl, S. Haridi, and K. Tzoumas. Apache Flink: Stream and batch processing in a single engine. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 36(4), 2015.

[17] M. Cemri, S. Agrawal, A. Gupta, S. Liu, A. Cheng, Q. Mang, A. Naren, L. E. Erdogan, K. Sen, M. Zaharia, A. Dimakis, and I. Stoica. AdaEvolve: Adaptive LLM driven zeroth-order optimization. *arXiv preprint arXiv:2602.20133*, 2026.

[18] A. C. Challenge. Overview. https://aicyberchallenge.com/overview/, 2025. Accessed: 2026-02-01.

[19] A. Cheng, S. Liu, M. Pan, Z. Li, S. Agarwal, M. Cemri, B. Wang, A. Krentsel, T. Xia, J. Park, S. Yang, J. Chen, L. Agrawal, A. Naren, S. Li, R. Ma, A. Desai, J. Xing, K. Sen, M. Zaharia, and I. Stoica. Let the barbarians in: How AI can accelerate systems performance research. *arXiv preprint arXiv:2512.14806*, 2025.

[20] W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, B. Zhu, H. Zhang, M. Jordan, J. E. Gonzalez, et al. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Forty-first International Conference on Machine Learning*, 2024.

[21] D. Chou and M. Jiang. A survey on data-driven network intrusion detection. *ACM Comput. Surv.*, 54(9), Oct. 2021.

[22] D. D. Clark. The design philosophy of the DARPA internet protocols. *SIGCOMM Comput. Commun. Rev.*, page 106–114, 1988.

[23] M. Corporation. MITRE ATT&CK® Framework. https://attack.mitre.org/.

[24] G. Deng, Y. Liu, V. Mayoral-Vilches, P. Liu, Y. Li, Y. Xu, T. Zhang, Y. Liu, M. Pinzger, and S. Rass. PentestGPT: Evaluating and harnessing large language models for automated penetration testing. In *Proceedings of the USENIX Security Symposium*, 2024.

[25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009.

[26] D. E. Denning. An intrusion-detection model. *IEEE Transactions on Software Engineering*, 13(2), 1987.

[27] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan, and R. Ranjan. Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Comput. Surv.*, 55(9), Jan. 2023.

[28] R. Elgedawy, J. Sadik, S. Dutta, A. Gautam, K. Georgiou, F. Gholamrezae, F. Ji, K. Lim, Q. Liu, and S. Ruoti. Occasionally secure: A comparative analysis of code generation assistants. *arXiv preprint arXiv:2402.00689*, 2024.

[29] D. Engler, D. Y. Chen, S. Hallem, A. Chou, and B. Chelf. Bugs as deviant behavior: A general approach to inferring errors in systems code. *SIGOPS Oper. Syst. Rev.*, 35(5):57–72, Oct. 2001.

[30] O. Etzion, M. Chandy, R. v. Ammon, and R. Schulte. Event-driven architectures and complex event processing. In *Proceedings of 2006 IEEE International Conference on Services Computing (SCC'06)*, 2006.

[31] R. Fang, R. Bindu, A. Gupta, and D. Kang. LLM agents can autonomously exploit one-day vulnerabilities. *arXiv:2404.08144*, 2024.

[32] N. Feamster, J. Rexford, and E. Zegura. The road to SDN: An intellectual history of programmable networks. *ACM SIGCOMM Computer Communication Review*, 44(2):87–98, 2014.

[33] K. J. Ferguson-Walter, M. M. Major, C. K. Johnson, and D. H. Muhleman. Examining the efficacy of decoy-based and psychological cyber deception. In *Proceedings of the USENIX Security Symposium*, 2021.

[34] S. Fort. AI found 12 of 12 OpenSSL zero-days (while curl cancelled its bug bounty). https://www.lesswrong.com/posts/7aJwgbMEiKq5egQbd/ai-found-12-of-12-openssl-zero-days-while-curl-cancelled-its, Jan. 2026. Accessed: 2026-02-01.

[35] R. Gennaro. Verifiable outsourced computation: A survey. In *Proceedings of the ACM Symposium on Principles of Distributed Computing*, 2017.

[36] V. Gligor. Security of emergent properties in ad-hoc networks. In *Proc. International Workshop on Security Protocols*, 2004.

[37] V. D. Gligor and S. I. Gavrila. Application-oriented security policies and their composition (position paper). In *Proceedings of the 6th International Workshop on Security Protocols*, 1998.

[38] V. D. Gligor and J. M. Wing. Towards a theory of trust in networks of humans and computers. In *Security Protocols XIX - 19th International Workshop, Cambridge, UK, March 28-30, 2011, Revised Selected Papers*, volume 7114 of *Lecture Notes in Computer Science*, 2011.

[39] A. Happe and J. Cito. Can LLMs hack enterprise networks? Autonomous assumed breach penetration-testing Active Directory networks. *ACM TOSEM*, 2025.

[40] M. A. Harrison, W. L. Ruzzo, and J. D. Ullman. Protection in operating systems. *Commun. ACM*, 19(8), Aug. 1976.

[41] V. Harsh, S. Sinha, H. Kamarthi, H. Milner, B. A. Prakash, V. Sekar, and H. Zhang. MoCE: A mixture-of-context aware experts framework for troubleshooting internet-scale services. In *Proceedings of the USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, 2026.

[42] H. He, C. Miller, S. Agarwal, C. Kästner, and B. Vasilescu. Speed at the cost of quality: How Cursor AI increases short-term velocity and long-term complexity in open-source projects. In *Proceedings of the International Conference on Mining Software Repositories*, 2026.

[43] J. L. Henning. SPEC CPU2006 benchmark descriptions. *ACM SIGARCH Computer Architecture News*, 2006.

[44] X. Hu, S. Li, T. Huang, B. Tang, R. Huai, and L. Chen. How simulation helps autonomous driving: A survey of Sim2Real, digital twins, and parallel intelligence. *arXiv preprint arXiv:2305.01263*, 2023.

[45] J. Huang and Q. Zhu. PenHeal: A Two-Stage LLM Framework for Automated Pentesting and Optimal Remediation. In *Proceedings of the Workshop on Autonomous Cybersecurity*. Association for Computing Machinery, 2024.

[46] IBM. SysFlow. https://sysflow.io/, 2025.

[47] K. Jesse, T. Ahmed, P. Devanbu, and E. Morgan. Large language models and simple, stupid bugs. In *Proceedings of the IEEE/ACM 20th International Conference on Mining Software Repositories (MSR)*, 2023.

[48] C. E. Jimenez, J. Yang, A. Wettig, S. Yao, K. Pei, O. Press, and K. Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? *arXiv preprint arXiv:2310.06770*, 2023.

[49] M. Kouremetis, D. Lawrence, R. Alford, Z. Cheuvront, D. Davila, B. Geyer, T. Haigh, E. Michalak, R. Murphy, and G. Russo. Mirage: Cyber deception against autonomous cyber attacks in emulation and simulation. *Annals of Telecommunications*, 2024.

[50] J. Kraprayoon, S. Ee, B. Rosen, Y. Matthew, A. Singh, C. Covino, and A. B. Gershovich. Highly autonomous cyber-capable agents: Anticipating capabilities, tactics, and strategic implications. *arXiv preprint arXiv:2603.11528*, 2026.

[51] B. W. Lampson. Hints for computer system design. In *Proceedings of the ACM Symposium on Operating Systems Principles*, page 33–48, 1983.

[52] S. Liu, S. Ponnapalli, S. Shankar, S. Zeighami, A. Zhu, S. Agarwal, R. Chen, S. Suwito, S. Yuan, I. Stoica, M. Zaharia, A. Cheung, N. Crooks, J. E. Gonzalez, and A. G. Parameswaran. Supporting our AI overlords: Redesigning data systems to be agent-first. *arXiv preprint arXiv:2509.00997*, 2025.

[53] Y. Liu, Y. Jia, R. Geng, J. Jia, and N. Z. Gong. Formalizing and benchmarking prompt injection attacks and defenses. In *Proceedings of the USENIX Security Symposium*, 2024.

[54] F. Long and M. Rinard. Automatic patch generation by learning correct code. *SIGPLAN Not.*, 51(1):298–312, Jan. 2016.

[55] F. Long and M. Rinard. Automatic patch generation by learning correct code. In *Proceedings of the ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, New York, NY, USA, 2016.

[56] R. Lopopolo. Harness engineering: leveraging codex in an agent-first world. https://openai.com/index/harness-engineering/, Feb. 2026.

[57] S. Maharjan, Q. Zhu, Y. Zhang, S. Gjessing, and T. Basar. Dependable demand response management in the smart grid: A Stackelberg game approach. *IEEE Transactions on Smart Grid*, 2013.

[58] M. L. Mazurek, S. Komanduri, T. Vidas, L. Bauer, N. Christin, L. F. Cranor, P. G. Kelley, R. Shay, and B. Ur. Measuring password guessability for an entire university. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*, CCS '13, page 173–186, New York, NY, USA, 2013. Association for Computing Machinery.

[59] W. Melicher, B. Ur, S. M. Segreti, S. Komanduri, L. Bauer, N. Christin, and L. F. Cranor. Fast, lean, and accurate: Modeling password guessability using neural networks. In *Proceedings of the USENIX Security Symposium*, Aug. 2016.

[60] H. Milner, Y. Cheng, J. Zhan, H. Zhang, V. Sekar, J. Jiang, and I. Stoica. Raising the level of abstraction for time-state analytics with the timeline framework. In *Conference on Innovative Data Systems Research (CIDR)*, 2023.

[61] A. Mirza, N. Alampara, S. Kunchapu, M. Ríos-García, B. Emoekabu, A. Krishnan, T. Gupta, M. Schilling-Wilhelmi, M. Okereke, A. Aneesh, et al. A framework for evaluating the chemical knowledge and reasoning abilities of large language models against the expertise of chemists. *Nature Chemistry*, 2025.

[62] A. Nadeem, D. Vos, C. Cao, L. Pajola, S. Dieck, R. Baumgartner, and S. Verwer. SoK: Explainable machine learning for computer security applications. In *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*, 2023.

[63] National Highway Traffic Safety Administration (NHTSA). Third amended standing general order 2021-01: Incident reporting for automated driving systems (ADS) and Level 2 advanced driver assistance systems (ADAS). NHTSA Standing General Order, June 2025. Effective Date: June 16, 2025.

[64] S. Oesch, A. Chaulagain, B. Weber, M. Dixson, A. Sadovnik, B. Roberson, C. Watson, and P. Austria. Towards a high fidelity training environment for autonomous cyber defense agents. In *Proceedings of the Cyber Security Experimentation and Test Workshop*, CSET '24, 2024.

[65] OpenAI. OpenAI o1 System Card. https://openai.com/index/openai-o1-system-card/, 2024.

[66] F. P. Osinga. *Science, Strategy and War: The Strategic Theory of John Boyd*. Routledge, 2007.

[67] B. Parno, A. Perrig, and V. Gligor. Distributed detection of node replication attacks in sensor networks. In *Proc. IEEE Symposium on Security and Privacy*, 2005.

[68] H. A. Pearce, B. Ahmad, B. Tan, B. Dolan-Gavitt, and R. Karri. Asleep at the keyboard? Assessing the security of GitHub Copilot's code contributions. In *Proceedings of the IEEE Symposium on Security and Privacy*, 2021.

[69] N. Perry, M. Srivastava, D. Kumar, and D. Boneh. Do users write more insecure code with AI assistants? In *Proceedings of the ACM Conference on Computer and Communications Security*, CCS '23, Nov. 2023.

[70] L. Phan, A. Gatti, Z. Han, N. Li, J. Hu, H. Zhang, C. B. C. Zhang, M. Shaaban, J. Ling, S. Shi, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

[71] M. Poess and C. Floyd. New TPC benchmarks for decision support and web commerce. *ACM SIGMOD Record*, 2000.

[72] Y. Potter, W. Guo, Z. Wang, T. Shi, H. Li, A. Zhang, P. G. Kelley, K. Thomas, and D. Song. Frontier AI's impact on the cybersecurity landscape. *arXiv preprint arXiv:2504.05408v4*, 2025.

[73] Red Hat. Ansible. https://www.ansible.com/, 2025.

[74] J. H. Saltzer, D. P. Reed, and D. D. Clark. End-to-end arguments in system design. *ACM Transactions on Computer Systems (TOCS)*, 2(4), 1984.

[75] F. B. Schneider. Least privilege and more [computer security]. *IEEE Security & Privacy*, 1(5):55–59, 2003.

[76] M. Sharif, P. Datta, A. Riddle, K. Westfall, A. Bates, V. Ganti, M. Lentzk, and D. Ott. DrSec: Flexible distributed representations for efficient endpoint security. In *Proceedings of the IEEE Symposium on Security and Privacy*, 2024.

[77] B. Singer, K. Lucas, L. Adiga, M. Jain, L. Bauer, and V. Sekar. On the feasibility of using LLMs to autonomously execute multi-host network attacks. *arXiv preprint arXiv:2501.16466*, 2025.

[78] B. Singer, Y. Saquib, L. Bauer, and V. Sekar. Perry: A high-level framework for accelerating cyber deception experimentation. In *Proceedings of the International Symposium on Research in Attacks, Intrusions and Defenses (RAID)*, 2025.

[79] S. Singh, Y. Nan, A. Wang, D. D'Souza, S. Kapoor, A. Üstün, S. Koyejo, Y. Deng, S. Longpre, N. A. Smith, B. Ermis, M. Fadaee, and S. Hooker. The leaderboard illusion. *arXiv preprint arXiv:2504.20879*, 2025.

[80] Splunk. Splunk SOAR. https://www.splunk.com/en_us/products/splunk-security-orchestration-and-automation.html, 2025.

[81] M. Standen, M. Lucas, D. Bowman, T. J. Richer, J. Kim, and D. Marriott. CybORG: A gym for the development of autonomous cyber agents. *arXiv preprint arXiv:2108.09118*, 2021.

[82] M. J. M. Turcotte, A. D. Kent, and C. Hash. *Unified Host and Network Data Set*, chapter 1, pages 1–22. World Scientific, 2018.

[83] S. Ullah, M. Han, S. Pujar, H. Pearce, A. Coskun, and G. Stringhini. LLMs cannot reliably identify and reason about security vulnerabilities (yet?): A comprehensive evaluation, framework, and benchmarks. In *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, 2024.

[84] United Nations. Digital public goods: Promoting open-source solutions for a more equitable world. https://www.un.org/digital-emerging-technologies/content/digital-public-goods.

[85] C. Wang and Z. Lu. Cyber deception: Overview and the road ahead. *IEEE Security & Privacy*, 16(2):80–85, 2018.

[86] Y. Wang, Q. Men, Y. Chen, J. Liu, G. Chen, Y. Zhang, G. Liu, and V. Sekar. Heimdall: Towards risk-aware network management outsourcing. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, 2025.

[87] A. Wilf, P. Aggarwal, B. Parno, D. Fried, L.-P. Morency, P. P. Liang, and S. Welleck. Propose, solve, verify: Self-play through formal verification. *arXiv preprint arXiv:2512.18160*, 2025.

[88] C. Yang, N. Neamtu, C. Hawblitzel, J. R. Lorch, and S. Lu. Verusage: A study of agent-based verification for rust systems. *arXiv preprint arXiv:2512.18436*, 2025.

[89] T. Yu, V. Sekar, S. Seshan, Y. Agarwal, and C. Xu. Handling a trillion (unfixable) flaws on a billion devices: Rethinking network security for the internet of things. In *Proceedings of the ACM Workshop on Hot Topics in Networks*, 2015.

[90] M. A. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica. Spark: Cluster computing with working sets. In *Proceedings of the USENIX Workshop on Hot Topics in Cloud Computing*, 2010.

[91] A. K. Zhang, N. Perry, R. Dulepet, J. Ji, C. Menders, J. W. Lin, E. Jones, G. Hussein, S. Liu, D. J. Jasper, P. Peetathawatchai, A. Glenn, V. Sivashankar, D. Zamoshchin, L. Glikbarg, D. Askaryar, H. Yang, A. Zhang, R. Alluri, N. Tran, R. Sangpisit, K. O. Oseleononmen, D. Boneh, D. E. Ho, and P. Liang. Cybench: A framework for evaluating cybersecurity capabilities and risks of language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[92] C. Zhang, Y. Park, F. Fleischer, Y.-F. Fu, J. Kim, D. Kim, Y. Kim, Q. Xu, A. Chin, Z. Sheng, H. Zhao, B. J. Lee, J. Wang, M. Pelican, D. J. Musliner, J. Huang, J. Silliman, M. Mcdaniel, J. Casavant, I. Goldthwaite, N. Vidovich, M. Lehman, and T. Kim. SoK: DARPA's AI cyber challenge (AIxCC): Competition design, architectures, and lessons learned. *arXiv preprint arXiv:2602.07666*, 2026.

[93] R. Zhang, C. Liu, W. Yu, M. Xu, Z. Wang, C. Su, X. Kang, Z. Yu, Y. Tang, et al. RCACopilot: On-call assistance via root cause analysis copilot. *arXiv preprint arXiv:2305.15778*, 2023.

[94] A. Zhong, D. Mo, G. Liu, J. Liu, Q. Lu, Q. Zhou, J. Wu, Q. Li, and Q. Wen. LogParser-LLM: Advancing efficient log parsing with large language models. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, 2024.

[95] L. Zhong and Z. Wang. Can LLM replace stack overflow? A study on robustness and reliability of large language model code generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[96] Y. Zhu, T. Jin, Y. Pruksachatkun, A. Zhang, S. Liu, S. Cui, S. Kapoor, S. Longpre, K. Meng, R. Weiss, F. Barez, R. Gupta, J. Dhamala, J. Merizian, M. Giulianelli, H. Coppock, C. Ududec, J. Sekhon, J. Steinhardt, A. Kellermann, S. Schwettmann, M. Zaharia, I. Stoica, P. Liang, and D. Kang. Establishing best practices for building rigorous agentic benchmarks. *arXiv preprint arXiv:2507.02825*, 2025.