Carnegie Mellon University
Software Engineering Institute

Toward the Use of Artificial Intelligence (AI) for Advanced Persistent Threat Detection

Matthew Walsh
Clarence Worrell
Thomas Scanlon

August 2024

TECHNICAL REPORT
CMU/SEI-2024-TR-001
DOI: 10.1184/R1/25282333

CERT Division

http://www.sei.cmu.edu

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

The view, opinions, and/or findings contained in this material are those of the author(s) and should not be construed as an official Government position, policy, or decision, unless designated by other documentation.

References herein to any specific entity, product, process, or service by trade name, trade mark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by Carnegie Mellon University or its Software Engineering Institute nor of Carnegie Mellon University - Software Engineering Institute by any such named or represented entity.

NO WARRANTY. THIS CARNEGIE MELLON UNIVERSITY AND SOFTWARE ENGINEERING INSTITUTE MATERIAL IS FURNISHED ON AN “AS-IS” BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.

DM24-0314

Table of Contents

List of Figures

List of Tables

Executive Summary

This report examines the feasibility and usefulness of implementing artificial intelligence (AI) and machine learning (ML) in cyber defense with a particular focus on advanced persistent threats (APTs).[1] APTs are cyber attacks carried out by well-resourced and sophisticated adversaries who target organizations to gain strategic advantage by exfiltrating data or disrupting operations. APTs leverage new and existing vulnerabilities in new and unexpected ways, seek to avoid detection, and adapt to defenders' actions. In particular, APTs typically generate very few signals for command-and-control systems unless they are activated for an attack. They can remain dormant for extended periods of time (e.g., years) thereby appearing to be part of normal operations when detectors are looking for operational changes. APT activation is a rare event, so APTs can escape typical detectors looking for repeating patterns. Thus, due to the stealthy and evolving nature of APTs, traditional tools for detecting and mitigating these types of cyber attacks may not be sufficient.

Over the past 10 years, there have been tremendous advances in AI, especially in deep learning, which is a type of ML algorithm that uses multiple-layer artificial neural networks (ANNs) to progressively extract higher level features from raw inputs. Recent applications of deep learning include computer vision, natural language processing, and Markov decision problems (e.g., Go, chess). These advances—along with other developments in generative AI, large language models, and game theory—have allowed organizations to apply AI to virtually all business functions, including cybersecurity.

Given the widespread adoption of AI by organizations and the explosion of academic research in this area, it is natural to consider the role of AI in APT defense. In this report, we examine the current state of AI-enabled APT defense. We begin by describing the stages that an APT must go through to succeed. Next, we perform a commercial market analysis of APT defenses. We then perform a bibliometric analysis to map out the academic research landscape on APTs. We identify three distinct challenges in APT defense and discuss how AI research addresses these challenges. We highlight the strengths and limitations of research on the use of AI for APT defense. Finally, we offer practical recommendations that will help organizations start incorporating AI into their layered APT defense strategies.

Our research reveals several key findings:

- There is evidence of significant APT attacks, making the threat operational and not theoretical.
- Many organizations offer AI-enabled solutions as part of a layered defense strategy that includes traditional methods (e.g., allowlists, denylists, signature matching, and firewalls). Traditional techniques are effective for detecting known threats, whereas AI methods are needed to identify novel ones, including APTs.

[1] Unless otherwise specified, our use of the term *AI* includes ML.

- Since 2010, the annual number of academic publications that cover cybersecurity has grown exponentially. Sub-areas that cover APTs and AI have also seen near-exponential growth. Hence, significant capabilities have yet to be transitioned from the literature to practical deployment.
- The results from academic research are promising, but several challenges must be overcome to transition this research into practice:
  - In APT detection, these challenges include minimizing false positives and ensuring the robustness of methods when facing new threats. There is expanding literature on the detection of rare events that could be integrated into detection tools to address some of these shortcomings.
  - In establishing a cyber-defense posture, these challenges include anticipating potential attacker actions and scaling algorithms for deployment that address real-world problems.
  - Two final challenges that cut across both areas are (1) developing testing algorithms on representative data sets or in representative simulation environments and (2) adopting open practices to improve research replicability.
- Organizations should include AI as part of a layered APT defense strategy. Traditional methods are insufficient given the novel and stealthy nature of APTs. Exploring the application of AI to APT defense can provide valuable insights into an organization’s cyber defenses and vulnerabilities, even if the AI solutions are not ultimately deployed.

Abstract

This report examines the feasibility and usefulness of implementing artificial intelligence (AI) and machine learning (ML) in cyber defense with a particular focus on advanced persistent threats (APTs). In this report, we examine the current state of AI-enabled APT defense. We begin by describing the stages that an APT must go through to succeed. Next, we perform a commercial market analysis of APT defenses. We then perform a bibliometric analysis to map out the academic research landscape on APTs. We highlight the strengths and limitations of research on the use of AI for APT defense. Finally, we offer practical recommendations that will help organizations start incorporating AI into their layered APT defense strategies.

1 Introduction and Overview

Advanced persistent threats (APTs) have become increasingly prevalent over the past 10 years. Unlike other cyber threats, APTs are orchestrated by well-resourced and sophisticated adversaries. They evolve over long periods of time, often remaining undetected for months or even years before acting on their final objective. Given the resources needed to conduct an APT, their targets are typically high value, and they can inflict great damage on their victims. These are a few examples:

- The Stuxnet APT emerged in 2009 and targeted industrial controllers, damaging physical equipment [Langner 2011]. By September 2010, Stuxnet infected approximately 100,000 hosts, most of which were located in Iran.
- The Carbanak APT emerged in 2013 and targeted financial institutions, causing nearly one billion dollars of cumulative losses to banks by 2016 [Johnson 2016]. The attackers conducted surveillance, including collecting video footage taken from employees' computers, to enable them to tailor their strategies to each specific bank's operational practices and vulnerabilities.
- The SolarWinds APT was initiated in 2019 and went undiscovered for fourteen months [GAO 2021]. To conduct the attack, hackers inserted malicious code into Orion, a network management system used to manage information technology (IT) resources. By inserting an exploit early in the supply chain, attackers compromised the data, networks, and systems of tens of thousands of organizations.

The MITRE Adversarial Tactics, Techniques, and Common Knowledge (ATT&CK) website identifies organized APT groups; the most alarming ones may be state sponsored [MITRE 2024]. There are many APT groups that have targeted critical infrastructure, defense organizations, government entities, and industrial bases. These APT groups can attack across multiple industry verticals, such as aerospace and aviation, satellite telecommunication, global positioning systems, industrial control systems, and manufacturing. Even closed networks and low connectivity systems can be vulnerable. Notably, Stuxnet could infect local computer networks by using Universal Serial Bus (USB) sticks [Langner 2011].

APTs are qualitatively different from other types of cyber threats. The National Institute of Standards and Technology (NIST) describes APTs in the following way [NIST 2011]:

> *The advanced persistent threat: (i) pursues its objectives repeatedly over an extended period of time; (ii) adapts to defenders' efforts to resist it; and (iii) is determined to maintain the level of interaction needed to execute its objectives.*

Stuxnet, Carbanak, and SolarWinds are examples of APTs that align with the NIST definitions of an APT. They were carried out over extended periods of time in a covert manner to avoid detection. Further, they sought to disrupt operations and exfiltrate data for strategic gain.

In this report, we illustrate how organizations can safeguard themselves against APTs with a particular emphasis on artificial intelligence (AI) and machine learning (ML).[2] It is important to consider AI as a defense because threat actors can use it as an offense. For example, attackers might use AI to do the following:

- obscure their presence to learn and mimic normal network behavior
- obscure data exfiltration by mimicking normal traffic
- improve attack-related scanning and search tasks
- spoof biometric-based authentication (e.g., voice, facial, fingerprint)
- detect countermeasures like honeypots
- use related techniques (e.g., game theory) to dynamically evade defensive measures

Defenders of high-value networks can be at a significant disadvantage when they do not leverage the AI technology that their attackers might use.

Academic research conducted about AI for cybersecurity has grown enormously. We review this literature, evaluate the feasibility of applying AI to APT defense, and provide practical guidelines that organizations can use to implement research findings in their operations.

1.1 Rise in Cyber Attacks

Over the last five years, there has been a steady increase in the number of significant cyber events. According to the Center for International and Security Studies at Maryland (CISSM) Cyber Events Database, as shown in Figure 1, the annual number of cyber events rose by more than 200 percent from 814 in 2014 to 1,918 in 2022 [Harry 2018]. This number of events includes only reported events and does not account for unsuccessful or unreported attacks, which may represent a significant proportion of the total number of attacks. Given the difficulty in detecting APTs, it is reasonable to assume that the growth in APTs is as least as fast as this data indicates.

[2] Unless otherwise indicated, our use of the term *AI* includes ML.

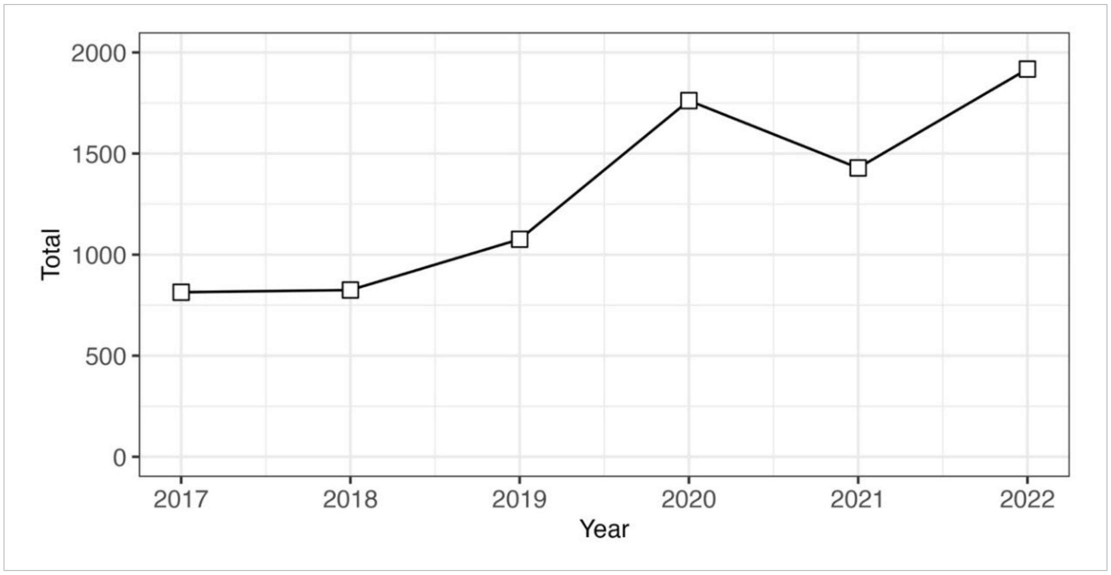

Figure 1: Annual Number of Significant Cyber Events Contained in the CISSM Cyber Events Database

To counter the growing threat of cyber attacks, organizations have taken three main approaches:

- They invest in Security Event and Incident Management (SEIM) cybersecurity measures. According to Statista's forecast for 2030, the global APT detection market size is expected to reach $657 billion by 2030, reflecting a 13 percent compound annual growth from 2019 to 2030 [Statista 2024].
- They purchase cyber insurance to recover any financial losses incurred from successful attacks [Marotta 2017].
- They apply zero trust architectures to contain the damage incurred following successful attacks. This strategy is paramount in a national defense scenario where monetary damage is incalculable (and thus uninsurable), but complete detection or prevention is unachievable.

While organizations recognize the importance of mitigating cyber risks, they also acknowledge that eliminating these risks may not be possible, so they employ some combination of these techniques.

1.2 Advances in AI and ML

Given the complexity of APTs and the overwhelming number of cyber events that must be monitored, it is prudent to explore the potential role of AI and ML in APT defense. AI involves machines performing tasks that are typically associated with human intelligence [Russell 2021]. The current wave of AI applications largely uses ML. ML refers to algorithms that learn to perform tasks based on historical data or interactions with the environment rather than relying on programmed rules [Jordan 2015]. ML is particularly valuable when humans do not know the optimal way to perform a task.

Over the last five years, academic work on AI for cyber defense has exploded. As shown in the top section of Figure 2, the annual number of publications in academic journals and conferences on this topic rose from 513 in 2015 to 6,021 in 2022. At the same time, as shown in the bottom of Figure 2, organizations have increasingly sought to leverage AI for many business functions, including cybersecurity [McKinsey 2022]. This demonstrates that AI has reached a level of maturity high enough to generate value for real-world applications.

Most recently, interest in generative AI has surged. Generative models can create convincing deepfake video, audio, and images. Large language models (LLMs) can generate complex text that appears to be written by a human, and these models excel at computer coding tasks such as generation, refactoring, and analysis. Generative AI tools are widely available, and often, using them does not require technical AI expertise. In this report, we focus on traditional AI and ML; however, in Section 6 (Conclusion) we identify the research needed to determine how generative AI might be used to defend against APTs as well as how adversaries might leverage generative AI when creating APTs.

Given the dual nature of using AI technology—for civilian and military applications—the core research used to detect APTs will likely come from the enormous investment driven by its use in areas other than APT detection. Future detection tools will likely result from mining this non-cyber defense research.

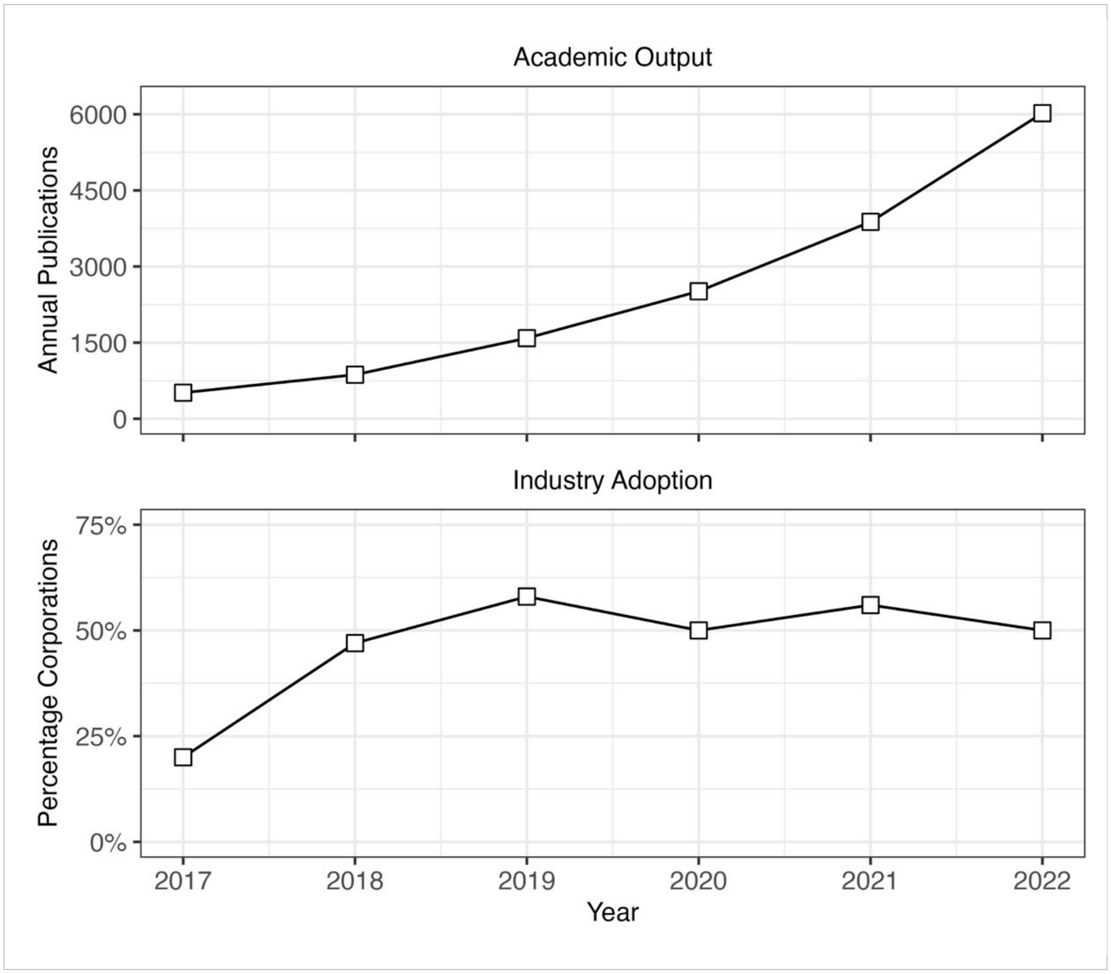

Figure 2: Annual Number of Academic Publications (Top) and Percentage of Corporations Reportedly Adopting AI (Bottom) from 2017 to 2022

1.3 Overview of This Report

In this report, we review the commercial markets and academic research on AI-enabled APT defense. Based on our findings, we conclude the following:

- AI offers significant potential to enhance traditional defenses and will play an essential role in APT protection, particularly in detecting zero-day exploits and other novel threats.
- More research is needed to transition theoretical AI defenses into practice. In particular, research publications frequently omit algorithm and data set details, making reproduction of results challenging. Transition requires additional research to fill in the missing information and validate capabilities against potential overstatements of results.
- Beyond ensuring that the science is adequate, transitioning research into practice identifies practical issues that can help make the science deployable. Known key challenges center on data-efficient learning, human-system integration, scalability, and test and evaluation in real-world settings.

We then offer practical steps that organizations can take to develop AI capabilities for APT defense. We conclude the following:

- Organizations can integrate AI-based methods into their APT defenses without disrupting current defenses and can gradually build an AI-based capability specific to the network(s) they protect.
- Implementing AI-based defenses can improve an organization’s awareness of its cyber-threat preparedness, yielding benefits regardless of whether AI solutions are ultimately deployed.

In the remaining sections of this report, we summarize the anatomy of an APT and general classes of APT defenses. We then provide a critical review of commercial APT defense products and academic research on APTs, focusing on AI-enabled solutions. Finally, we present a framework that organizations can use to begin incorporating AI into their APT defenses.

2 Anatomy of an APT

The term *advanced persistent threat* was used in 2011 to describe a new class of manual intrusions that surpassed automated viruses and worms [Hutchins 2011]. Since then, numerous federal and private organizations have sought to define the key phases, or *anatomy*, of an APT. In parallel, they have developed defenses to thwart APTs at each stage. This section provides a brief overview of these efforts.

2.1 Behavioral Model

The Cyber Kill Chain®, developed by Lockheed Martin, is a model that breaks down APT attacks into a sequence of steps [Lockheed Martin 2024]:

1. **Reconnaissance.** The attacker conducts research to identify potential vulnerabilities in the target's system.
2. **Weaponization.** The attacker combines malware with an exploit to create custom payloads.
3. **Delivery.** The attacker sends the weaponized payload to the target.
4. **Exploitation.** The attacker uses an exploit to gain access to the target system.
5. **Installation.** The attacker installs a persistent backdoor to maintain access to the compromised system.
6. **Command & Control.** The attacker uses malware to establish a channel for controlling the compromised system.
7. **Actions on Objectives.** The attacker achieves the goal of the attack.

The Cyber Kill Chain implies that each step must be carried out in sequence, and thus, disrupting any of the steps can cause the attack to fail. This model has become a widely accepted framework for understanding the anatomy of APT attacks.[3]

2.2 Defense Taxonomy

APT defenses can be divided into three categories: monitoring, detection, and mitigation [Alshamrani 2019].[4]

- **Monitoring** involves the ongoing collection of data at both the network and node levels to facilitate defense.

[3] There are other models of APT behavior, most of which include more granular descriptions of an adversary's actions once they have defeated perimeter defenses (i.e., during steps 6 and 7 of the Cyber Kill Chain®). Nonetheless, the Cyber Kill Chain describes APT attacks with enough detail to facilitate analysis while remaining simple enough to be clear and communicable.

[4] Subversion and other forms of deception are a special case of mitigation.

- **Detection** aims to distinguish between benign and malicious activity and works in tandem with monitoring. Anomaly detection and pattern matching are two general classes of ML methods used for detection. Anomaly detection analyzes node or network behavior to identify events that deviate from those generated by typical behavior. Pattern matching compares node or network events to known patterns of malicious activity.
- **Mitigation** refers to the prevention or reduction of attacks and can be proactive or reactive. Proactive methods actively disrupt an attacker's activities and include techniques such as deception and moving target defenses.[5] Reactive methods identify potential attack scenarios based on known vulnerabilities in the system and consider how an ongoing attack may unfold.

2.3 Defensive Methods with Respect to Cyber Kill Chain Steps

Defense techniques can be applied to different steps in the Cyber Kill Chain. Table 1 shows several examples of these intersections. For instance, an organization can monitor network traffic in the form of packets. By analyzing packet attributes (e.g., source Internet Protocol [IP] address, destination IP address, destination port, packet timing, packet size), the organization may determine that an attacker is scanning for vulnerable access points, such as open ports. They may then apply a predetermined, event-based moving target defense to change the attack surface, rendering the attacker's reconnaissance obsolete.

Table 1: Intersections Between Cyber Kill Chain Attack Stages and Defense Methods

Stage	Monitoring	Detection	Mitigation
Reconnaissance	Packet monitoring	Port scanning or network mapping	Use moving target defense to degrade attacker intelligence.
Weaponization	Threat intelligence reports	Network vulnerabilities	Develop targeted defenses against specific threats.
Delivery	Emails, altered control plane information, software updates	Malicious attachments or links	Allocate defense resources across highest value nodes.
Exploitation	Code monitoring	Malware indicators	Set policy for deciding which indicators to inspect manually.
Installation	Log monitoring	Unusual or unexpected activity, such as failed login attempts, system file changes, privilege escalation	Balance recovery actions (e.g., reinstallation) with service interruption costs.

[5] Honeypots are decoy assets intended to distract or discourage the attacker while possibly allowing the defender to gain insight into the attacker's tactics. Moving target defenses change network and/or host structures to undermine attacker reconnaissance and surveillance.

Stage	Monitoring	Detection	Mitigation
Command & Control	Packet and system operational monitoring	Pattern matching looking for "beaconing" between an internal host and external domain name, deviations from expected traffic patterns, deviations from expected operational parameters	Employ subversion to lure attackers into virtual traps.
Actions on Objectives	Log monitoring	Forensic analysis to trace attack	Use attack graphs to anticipate attacker moves.

When forming a monitoring strategy, organizations must decide which parts of their network and/or hosts they want to monitor, which data elements to collect, and how to store that data. These decisions entail a tradeoff between having a comprehensive record of events and minimizing the costs associated with capturing and storing data. Threat intelligence reports can be used to determine likely attack vectors, and organizations can use this information to identify associated threat indicators and the corresponding data elements necessary for generating those indicators. While threat intelligence reports allow organizations to prepare for threats that they have not directly experienced, they do not allow them to prepare for truly novel threats.

Detection requires defining the combinations of indicators—or *features*—that trigger threat warnings. This involves a tradeoff between limiting the number of benign events classified as malicious (false positives) while also limiting the number of malicious events classified as benign (false negatives). However, even with a low false positive rate, the number of alerts generated can still be overwhelming given the high volume of cyber events that occur daily.

Regarding mitigation, organizations must decide how to allocate finite defense resources. APT mitigation is costly and may disrupt service, so organizations must balance increased security with the costs and potential loss of service quality [CIS 2023].

Interrupting any stage in the Cyber Kill Chain can cause an attack to fail. Therefore, organizations usually adopt a combination of monitoring, detection, and mitigation strategies, as shown in Table 1. In addition to these strategies, organizations can also use security awareness training, patch management, firewalls, anti-virus software, content filtering, and other traditional measures to reduce risk.

2.4 Summary

APTs are multi-stage attacks. By disrupting any stage, defenders can defeat an attack. To do so, defenders must use monitoring, detection, and mitigation strategies. However, due to the complex attack surface and the vast number of daily cyber events, selecting effective and resource-efficient policies poses a challenge. In the following sections, we explore using AI in potential solutions that help defenders make better-informed decisions.

3 Commercial and Academic Landscape Analysis of Cyber Defense Products Using AI

As organizations consider augmenting their cyber defense postures with AI-enabled technologies, they must understand the current landscape of commercial products, some of which are dual use. Organizations must also understand the state of academic research and the future technologies that may become possible. To provide insight into these issues, we performed a landscape analysis of commercial offerings and academic research about using AI for APT defense.

3.1 Commercial Analysis

To identify industry leaders, technology trends, and market gaps in commercial APT defense products, we performed a market analysis focused specifically on using AI for APT threat detection and mitigation.

3.1.1 Methods

To make our analysis manageable, we focused on organizations included in one of the following *quadrant* reports for the APT defense market:

- Radicati: *APT Protection Market Quadrant 2021* [Radicati 2021]
- Forrester: *Now Tech: Enterprise Detection and Response, Q1 2020* [Zelonis 2020]

These reports identify top vendors in APT defense and categorize them into four quadrants based on strategic vision (narrow versus broad) and maturity (emerging versus established).

Using these reports and other sources, we identified 22 organizations that offer a range of commercial APT defense solutions, all of which advertise the use of AI or ML in one or more of their products (Table 2). We reviewed publicly available material from each organization to assess their APT defense offerings. Our analysis was limited by the fact that organizations guard their proprietary information, which prevented us from fully evaluating the technical details and performance characteristics of all solutions. Nonetheless, we were able to identify some of the types of functionality provided by each organization.

3.1.2 Results

Table 2 presents a summary of the APT defenses that each organization offers. The table also shows the AI techniques used in their products along with whether their solutions focus on threat monitoring, detection, or mitigation.

Table 2: Comparison of Commercial APT Defense Products

Organization	AI Methods Described	Monitoring	Detection	Mitigation
Vendor 1	Not identified	🟩	🟩	🟧
Vendor 2	Not identified	🟩	🟧	🟧
Vendor 3	General linear model, tree-based method, support vector machine	🟩	🟩	🟩
Vendor 4	Not identified	🟩	🟩	🟩
Vendor 5	Ensemble	🟩	🟩	🟧
Vendor 6	Not identified	🟩	🟩	🟧
Vendor 7	Tree-based methods	🟩	🟩	🟧
Vendor 8	Neural networks	🟩	🟩	🟧
Vendor 9	Not identified	🟩	🟩	🟧
Vendor 10	Not identified	🟩	🟩	🟧
Vendor 11	Not identified	🟩	🟩	🟧
Vendor 12	Neural networks, reinforcement learning	🟩	🟩	🟩
Vendor 13	Not identified	🟩	🟩	🟩
Vendor 14	Neural networks	🟩	🟩	🟧
Vendor 15	Neural networks	🟩	🟩	🟩
Vendor 16	Not identified	🟩	🟩	🟧
Vendor 17	Not identified	🟩	🟩	🟩
Vendor 18	Neural networks	🟩	🟩	🟧
Vendor 19	Not identified	🟩	🟩	🟩
Vendor 20	Probabilistic graphical models, hierarchical clustering, locality sensitive hashing	🟩	🟩	🟧
Vendor 21	Expert systems	🟩	🟩	🟧
Vendor 22	Neural networks	🟩	🟩	🟩

Green = Yes; Orange = No

As shown in Table 2, all products offered some form of threat detection (i.e., anomaly detection or pattern matching). These products primarily focused on threat detection, antivirus, event correlation, phishing email detection, blacklists, and vulnerability detection. All products also included some form of monitoring. Since detection methods are applied to data sources, monitoring is necessary to drive effective detection.

Only 8 of 22 organizations provided products for threat mitigation. Of these, 6 used graph analysis to trace threats within a network, 2 involved honeypot defenses, and 1 used both honeypot and moving target defenses.

All organizations claimed to use AI, yet few explained precisely how. As seen in Table 2, 6 organizations reported using neural networks, and 2 reported using tree-based algorithms. Each of the remaining methods (i.e., general linear models, support vector machines, ensembles, reinforcement learning, probabilistic graphical models, hierarchical clustering, and expert systems) was reported by only a single organization. Lastly, only 3 organizations reported performance benchmarks:

- Vendor 6 reported 99% recall.
- Vendor 16 reported 95% recall on new threats.
- Vendor 22 reported 99.98% accuracy.

These values are difficult to interpret without context. The more general point is that commercial APT defenses are not described in sufficient detail to predict their performance in new and different settings.

3.2 Bibliometric Analysis

Bibliometric analysis involves using quantitative techniques to analyze bibliometric data, such as academic publications and citations [Borgman 2005]. The goal of this analysis is to gain insight into the intellectual structure and emerging trends in a particular field of study. In this section, we report the results of our bibliometric analysis of APTs and the growing role of AI in this area.

3.2.1 Methods

We collected document data from Scopus, a database that contains approximately 90 million records from hundreds of thousands of curated journals and proceedings [Scopus 2024]. We limited our search to journal articles and conference papers from 2010 to 2022 that included variations of the phrases *cybersecurity* or *advanced persistent threat* in the title, keywords, or abstract. From these articles, we identified those that also included variations of the phrases *machine learning* or *artificial intelligence* in the title, keywords, or abstract.

3.2.2 Results

Figure 3 shows the annual growth of publications about cybersecurity (left) and APTs (right) from 2010 to 2022. During this time, the growth of publications in both areas was exponential.

Figure 3 also shows the annual growth of AI-focused publications on these topics from 2010 to 2022. In 2017, the total number of AI-focused publications about cybersecurity and APTs reached 207 and 18, respectively. Just five years later in 2022, the numbers reached 6,021 and 251. Thus, the number of new academic papers on cybersecurity and APTs each year has exploded, and ones that focused on AI made up a growing share.

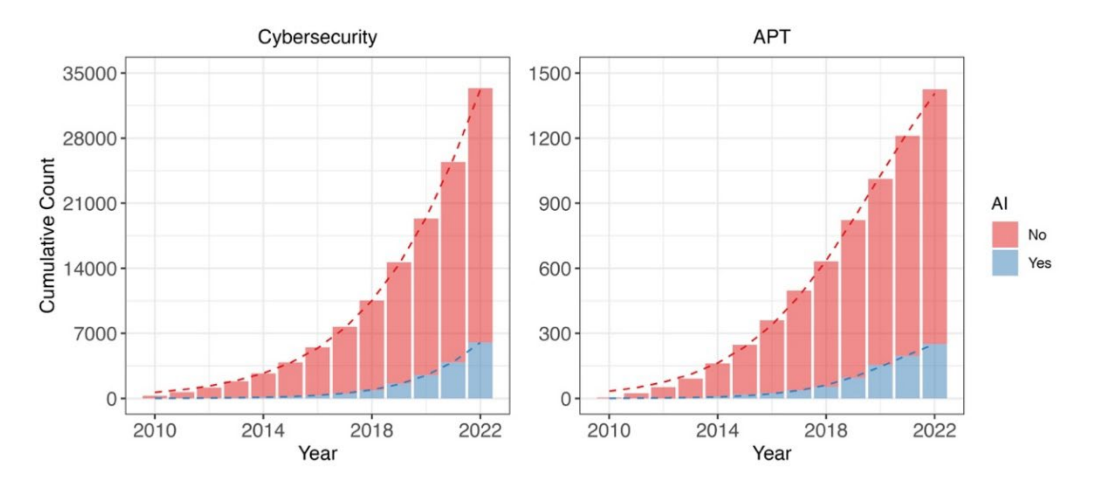

Figure 3: Cumulative Number of Indexed Publications About Cybersecurity and APT from 2010 to 2022

Next, we investigated the AI approaches explored in APT research. To ensure comprehensive coverage, we expanded our dataset by incorporating 114 additional records from Web of Science (WoS), a curated database of journal and conference papers that partially overlaps with Scopus [Clarivate 2024]. We then standardized the keywords and categorized them into seven classes of AI and ML techniques:

- graph model
- natural language processing
- optimization
- reinforcement learning
- unsupervised learning
- game theory
- supervised learning

For instance, we grouped keywords such as *support vector machine*, *gradient boosting*, and *random forest* under *supervised learning*, and we grouped *k-means* and *clustering* under *unsupervised learning*.

Figure 4 illustrates the number of publications that contain keywords that correspond to the seven classes of AI and ML techniques over the entire period. Because each publication could contain more than one keyword, we counted some articles in more than one category. Our analysis revealed that the largest percentage of publications focused on *supervised learning* (31.3 percent), followed by *game theory* (14.7 percent), and then by *unsupervised learning* (5.8 percent). As we discuss in Section 4, the emphasis on supervised learning approaches may be limiting given the need for labeled data to use these approaches. This data may not exist for APTs.

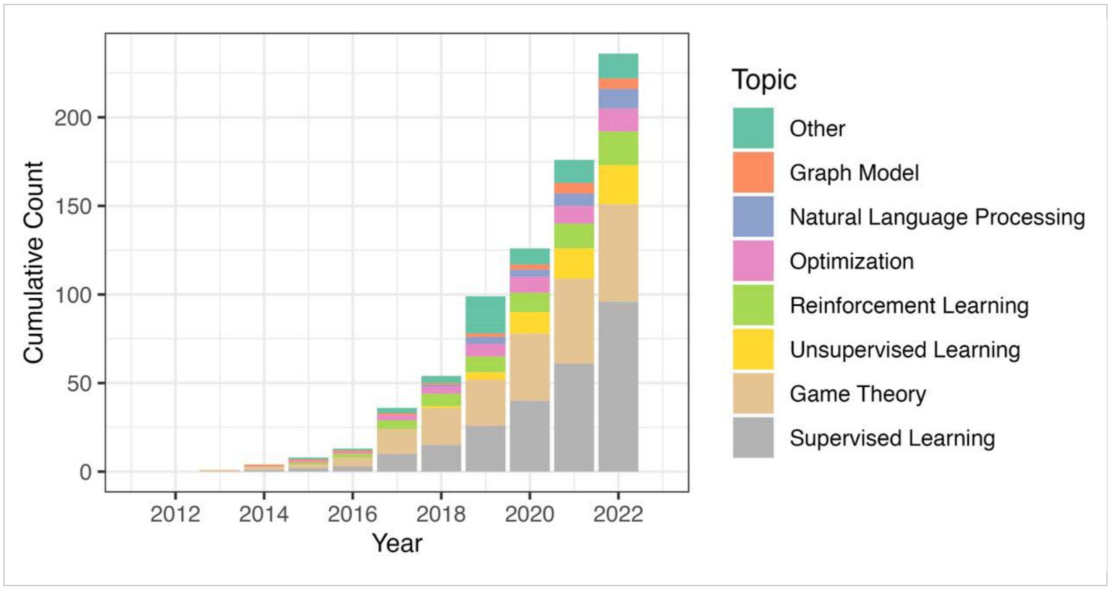

Figure 4: Total Number of Indexed Papers About APT by AI and ML Class from 2012 to 2022

3.3 Summary

In our commercial landscape analysis, we found that many organizations are currently developing and advertising (but not specifying) AI-enabled APT defenses. These include well-established organizations with broad offerings as well as those with a narrower focus. Among the organizations that describe aspects of their AI-enabled solutions, neural networks are the most common approach they used. However, these methods and performance characteristics are not described in enough detail to determine whether they will be effective against different threats or in different settings. As compared to traditional cyber-defense products, this concern is exacerbated for ML-enabled products due to their reliance on the data sets and environments used in training.

Many products that incorporate AI also include traditional cybersecurity techniques, such as allowlists/denylists, signature matching, and firewall rules. Traditional techniques are useful for detecting *known threats*, while AI and ML can help identify *novel ones*. The point is that commercial products are using AI to augment (not replace) other APT defenses.

In our bibliometric analysis, we found that the number of publications that cover cybersecurity in general, and AI specifically, has exploded since the early 2010s. In terms of the taxonomy of AI defenses described earlier, academic literature that covered *supervised* and *unsupervised learning* tends to map to *detection*, whereas the literature on *game theory* maps to *mitigation*. The difference in the proportional share of the number of papers on these topics suggests that AI use cases for *detection* are more mature than ones for *mitigation*.

Should organizations invest in the development and transition of AI-enabled APT defenses arising from academic research? In Section 4, we critically evaluate this academic body of work. To preview, we find that many technical and implementation challenges remain. Thus, as compared to commercial AI products, AI systems described in APT research have lower technology readiness levels. To transition these systems into use, organizations would first need to invest time and resources. However, in doing so, it could provide stronger assurances about the behaviors and performance of the resulting systems. Further, while developing systems, organizations could take steps to ensure that they adhere to guidelines for safe AI, as detailed in publications like NIST's *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* [NIST 2023].

4 Challenges in APT Defense

APTs pose at least three distinct challenges that have been partially addressed through academic research:

- **Challenge 1: Detecting Novel Attacks.** Detect intrusions that use zero-day exploits or that use known exploits in new ways. This is a *detection* problem that requires probabilistic judgments based on the features of cyber events, since each APT is unique and differs from past attacks.
- **Challenge 2: Low-and-Slow Movements of the Attacker.** Integrate low-level indicators into a comprehensive assessment of an APT. This is a *sensemaking* problem. APTs move slowly and attempt to evade detection. Methods for understanding cyber events must gather and combine information from multiple weak indicators.
- **Challenge 3: Defending a Complex Attack Surface.** Defend complex attack surfaces presented by large networks. This is a *decision* problem. APTs may target multiple nodes, systems, and vulnerabilities in a network, making it challenging to block, detect, and evict attackers. To do so effectively, defenders must deploy limited resources efficiently and consider the values of nodes; attackers' objectives; and their tactics, techniques, and procedures.[6]

In this section, we explore how AI research attempts to address these challenges.

4.1 Challenge 1: Detecting Novel Attacks

There are two basic types of intrusion detection systems (IDSs): signature-based and anomaly-based systems:

- **Signature-based IDSs** use a database of known attack patterns (i.e., signatures) that have occurred in the past. The primary strength of these methods is that they are effective at blocking known threats. Their primary weakness is that they may not detect novel threats.
- **Anomaly-based IDSs** involve constructing models of normal network traffic and using those models to detect when traffic is abnormal. The primary strength of these methods is that they do not depend on the signature of a particular type of *known* attack. The primary weakness of these methods is that not all anomalies are threats, and not all threats produce anomalies.

While neither type of IDS can detect all threats, both can detect some threats, which justifies their inclusion in an ensemble of defenses.

ML is a natural method for creating signature-based and anomaly-based IDSs. The complexity of cyber data is that human analysts may not know the optimal sets of rules to use for monitoring and sorting events. Given historic data, ML techniques can learn effective rules for detecting threats and identifying anomalies. There has been extensive research in this area [Bhuyan 2013].

[6] APTs can be introduced outside the target environment. For example, they can be introduced earlier in the supply chain. Thus, APT defenses must also be applied outside the target environment.

The resulting collection of ML-based IDSs can be categorized into two classes based on the type of learning they use: supervised and unsupervised, which we explore next.

4.1.1 Problem Definitions

In *supervised* learning, a model is trained to accurately predict the labels or values of future outcomes. For example, it may determine whether a cyber event is malicious or benign, as show on the left side of Figure 5. The model is trained using historic data, and each cyber event in the historic data is represented as a record with a set of input features. Examples of these features include source IP address, destination IP address, protocol, packet size, and packet timing (in the case of NetFlow data). Each record also has a label, such as whether it was normal or anomalous traffic.

Supervised learning trains a model to accurately predict the labels for each record given its corresponding input features. Once trained, the model can be used to classify new events as normal or anomalous.

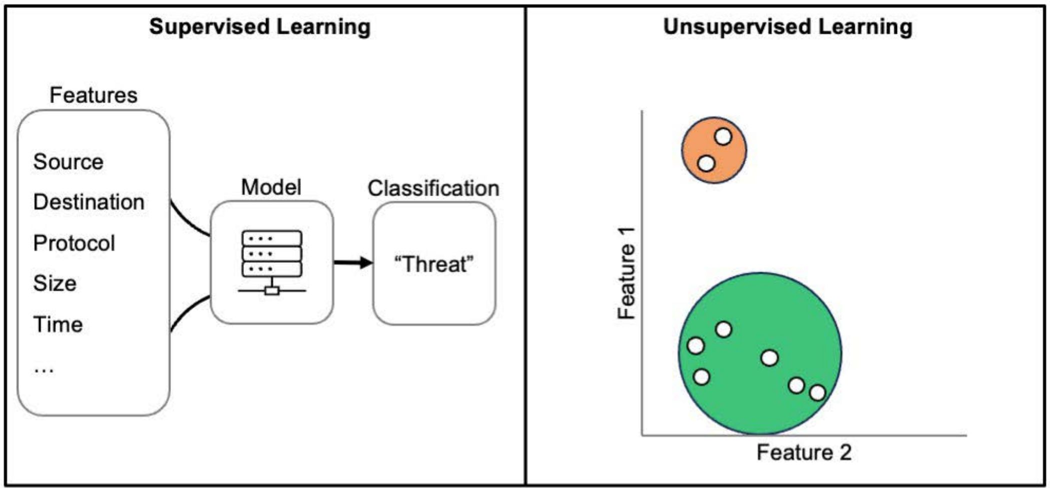

Figure 5: Supervised and Unsupervised Learning

In *unsupervised* learning, a model is trained from unlabeled data, as shown on the right side of Figure 5. Unlike supervised learning, which attempts to predict outcomes, unsupervised learning attempts to discover the structure of the data itself. One use of unsupervised learning in IDSs is to identify anomalies as events that are *distant* from most other events, which are assumed to be benign.

Choosing between supervised and unsupervised learning depends on the availability of labeled training data. If a large amount of correctly labeled data is available, supervised learning can be used. However, the application of supervised learning methods to APT detection presents the following three challenges:

1. Since most traffic is normal, historic data may contain few examples of attacks.
2. Since APTs seek to avoid detection, data labeled as benign may contain attacks.
3. Given their novel and evolving nature, new APTs may not resemble historic attacks.

Unsupervised learning does not require labeled training data, making these techniques complementary to supervised learning. A key assumption is that anomalies can be characterized by some measure of distance from normalcy (e.g., the Euclidean distance between two vectors of continuous variables). However, the application of unsupervised learning methods to APT detection presents the following two challenges:

1. Measures of distance are calculated with respect to the features provided. Given the high dimensionality of cyber data sources, domain knowledge—including assessment of potential adversary tactics to conceal actions—must be used to select only the most meaningful features.
2. Data may be anomalous in many ways. Thus, while unsupervised learning methods may discover an underlying structure, the structure may not separate benign from malicious events.

The implication is that to successfully apply supervised or unsupervised learning to IDSs, the methods must be tailored to the data and environment of the use case.

4.1.2 Academic Research on Anomaly-Based IDSs

Buczak and Guven extensively review intrusion detection methods that use ML [Buczak 2015]; they provide examples of signature-based and anomaly-based detection methods that use a variety of ML approaches: artificial neural networks, association rule mining, Bayesian networks, clustering, decision trees, ensemble learning, evolutionary computation, hidden Markov models, inductive learning, naïve Bayes, sequential pattern mining, and support vector machine.

The methods that Buczak and Guven describe in their paper, though developed for use in other domains and settings, are applicable to signature-based and anomaly-based detection. However, the authors note that, in tasks like detecting abnormalities in medical images, features and outcomes change very gradually. In cybersecurity, attacker behaviors may change daily, which has implications for the concurrency of data available to train models, the importance of monitoring model performance, and the frequency of model updates.

In 2017, Aburomman and Reaz survey the intrusion detection literature surrounding ensemble methods [Aburomman 2017]. There are many ML approaches that can be used for classification—categorizing records into separate groups (e.g., benign versus malicious events). The intuition behind an ensemble approach is that, although the predictions of different methods are correlated, they provide distinct information. Thus, the predictions of many weak classifiers can be aggregated so that they outperform any one classifier. Further, ensemble classifiers may be more difficult to defeat because they employ multiple methods with different vulnerabilities.

In 2020, Ferrag and their coauthors surveyed research on deep learning methods for intrusion detection [Ferrag 2020]. One of the key strengths of deep learning methods is their ability to learn complex relationships between input features and outcomes. Given the complexity of cyber data, this ability may allow deep learning methods to outperform other supervised learning approaches. In their review, Ferrag and their coauthors considered applications of recurrent neural networks, deep neural networks, restricted Boltzmann machines, deep belief networks, convolutional neural networks, deep Boltzmann machines, and deep autoencoders to intrusion detection. They found that these and other deep learning methods performed well.

4.1.3 Future Research Directions for Anomaly-Based IDSs

The significant and sustained interest in AI-based network intrusion detection methods, especially for anomaly detection, suggests that ML may ultimately play an important role in practical intrusion detection. However, several challenges must be addressed for this to occur. Table 3 summarizes these challenges along with potential solutions.

Table 3: Challenges and Solutions for Anomaly-Based IDSs

Challenge	Description	Solutions
Data efficiency	Although stores of cyber data may be vast, they can contain few examples of attacks. Additionally, due to the evolving nature of threats, historic data may quickly lose its concurrency.	• Use data-efficient learning methods. • Apply transfer learning. • Apply data-augmentation techniques. • Use Bayesian methods and other approaches that can leverage expert knowledge.
False alarms	False alarms refer to benign events classified as malicious. Given the high volume of information that cybersecurity teams must handle each day, even a system with a low false alarm rate can produce an unacceptable increase in operator workload.	• Work with frontline workers to set acceptable thresholds to balance false alarms and misses. • Improve algorithm performance. • Explore ways to combine detectors to reduce false alarms.
Model drift	Network traffic patterns naturally change over time. Additionally, attackers are constantly developing new methods to evade detection. Thus, the performance of detection methods will degrade over time.	• Continuously monitor environment drift. • Continuously monitor model performance. • Frequently retrain models. • Use methods that are robust against environment change.
Computational complexity	Some network defenses must be applied in real time. Additionally, these defenses may need to be applied by edge computing devices. The computational complexity, or time needed to run inference algorithms for detecting threats, may limit their applicability.	• Measure time complexity during testing. • Develop low-power and high-performance edge computing devices. • Explore options for distributed and cloud-based processing.
Explainability	Some ML methods trade off explainability for performance. To engender trust and enable verification, human operators must be able to understand how algorithms arrived at their decisions.	• Select inherently interpretable methods. • Integrate explanation interfaces with complex models.

Challenge	Description	Solutions
Benchmarking	Lack of agreed-on measures of performance (MoPs), measures of effectiveness (MoEs), and benchmark data sets, make it difficult to perform an “apples-to-apples” comparison of ML methods.	• Reach consensus in the cyber-defense community about relevant MoPs and MoEs. • Evaluate methods using those metrics. • Evaluate methods using a holdout data set and independent data set to assess generalizability.
Replicability	Replicability (i.e., the ability to reproduce another researcher’s work and arrive at the same conclusions) is a general challenge in science. This also extends to implementing ML systems.	• Adopt minimum reporting standards to ensure replicability of cyber ML results. • Make models and methods available in code repositories.

4.2 Challenge 2: Low-and-Slow Movements of the Attacker

Low-and-slow movements are the hallmark traits of an APT. The attacker makes strategic and incremental steps over a long period of time (sometimes years), taking time to plan the next move. Therefore, a central challenge in APT detection is recognizing a sequence of events occurring over a long period of time, when each individual event might not be alarming but the events in aggregate indicate a possible APT. A related challenge, “needle in a haystack,” is when the evidence of these steps can be masked by the very high volume of normal traffic.

Figure 6 shows a simplified version of the problem of correlating indicators and arriving at a high-level interpretation of events. The underlying state of the environment, which cannot be directly observed, indicates whether an APT is underway. Defenders observe streams of indicators, which provide weak and imperfect evidence of an APT. By integrating information from indicators across time, defenders can amplify weak signals, boosting detection. Integrating information can also mitigate noisy signals and reduce false alarms.

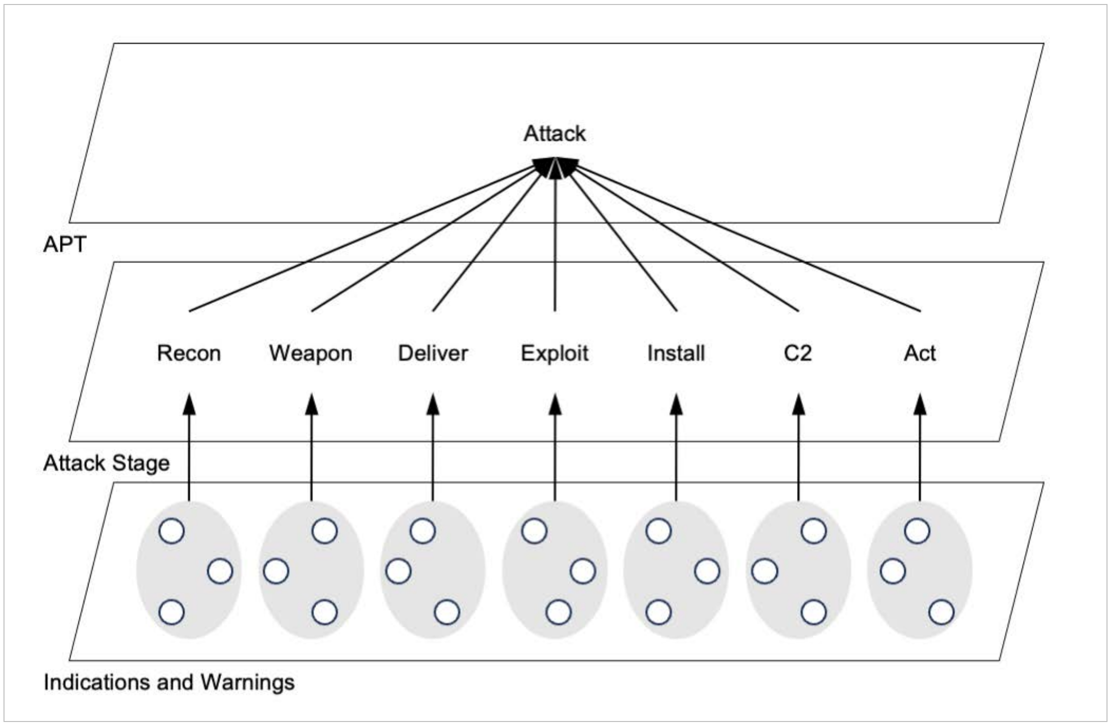

Figure 6: APT Evidence Accumulation

4.2.1 Academic Research on Event Integration

MLAPT [Ghafir 2018] and HOLMES [Milajerdi 2019] are two APT detection frameworks that involve integrating lower level alerts to identify sequences of events that might be part of the same APT.

MLAPT uses three main steps: threat detection, alert correlation, and attack prediction. In the *threat detection* step, network traffic is monitored, and alerts are generated for disguised executable file detection, malicious file hash detection, malicious domain name detection, malicious IP address detection, malicious Secure Sockets Layer (SSL) certificate detection, domain flux detection, scan detection, and Tor connection detection. In the *alert correlation* step, MLAPT clusters individual alerts so that each cluster contains events potentially generated by the same APT. Finally, in the *attack prediction* step, an ML model is used to predict the likelihood of events in each cluster evolving into a full APT attack, which provides network security team members with a mechanism to triage their work.

HOLMES starts with audit logs from multiple hosts and captures events relating to users, files, memory, processes, and network connections. Next, a provenance graph of the log data is constructed in which the vertices represent processes and objects, and the edges represent dependencies between processes and objects. HOLMES then can be used to map low-level audit log events through an intermediate layer of rules and up to high-level APT steps. Scenario graphs are constructed that contain events that have occurred in the correct sequences to indicate a possible

APT. Finally, a ranking scheme is applied to the scenario graphs to identify those most indicative of a possible APT.

Both MLAPT and HOLMES share the idea that they can integrate multiple low-level indicators into intermediate and high-level representations of APTs. This integration might reduce the total number of false alarms, amplify weak alerts associated with one another, and present interpretable evidence of APTs to human analysts. MLAPT and HOLMES are not far removed from work on automated indications and warnings because these frameworks integrate signals arising from these automated detectors.

4.2.2 Future Research Directions for Sensemaking

There has been far less research about automated sensemaking than AI-based intrusion detection. Currently, the primary challenge is to demonstrate feasible technical approaches for using AI to extract meaning from streams of event alerts and indicators. MLAPT and HOLMES were evaluated using a private dataset and another one provided by the Defense Advanced Research Projects Agency (DARPA) Transparent Computing Program. While the results of the evaluations are promising, further testing is needed. Table 4 summarizes the challenges for automated sensemaking along with potential solutions.

Table 4: Challenges and Solutions for Automated Sensemaking

Challenge	Description	Solutions
Effective technical approaches for information integration	Intrusion detection is essentially a supervised learning problem that can draw on well-established solutions from non-cyber domains. Bayesian and graph-based methods for sensemaking are less technically mature.	• Conduct literature reviews to discover promising methods from non-cyber domains. • Invest in developing Bayesian and graph-based methods for sensemaking.
Dependence on detectors	Alerts arising from lower level detectors provide inputs to sensemaking systems. System performance depends on the performance of low-level detectors.	• Invest in developing low-level detectors.
Passive nature of information searching	Systems have access to predetermined elements of information; whereas human analysts may actively seek different information.	• Invest in developing active sensing capabilities to allow systems to seek information and disambiguate scenarios.
Decoupling simultaneous attacks	In a large enterprise, multiple indications and warnings, though occurring in close temporal proximity, may nonetheless arise from different attackers.	• Test system experiences for multiple concurrent APTs.

4.3 Challenge 3: Defending a Complex Attack Surface

Many cybersecurity problems involve decision making in uncertain, multi-agent conditions. Game theory provides a set of mathematical tools for formalizing and informing these types of decisions [Kamhoua 2021]. The chief benefit of applying game theory to cybersecurity is that it enables *rational* decision making. Subject to the set of assumptions made about the problem and adversary, game theory provides provably optimal strategies.

Multi-agent reinforcement learning (MARL) is a sub-field of reinforcement learning used to study the behaviors of agents that coexist in shared environments. Each agent is motivated to maximize its rewards. MARL relates to game theory, but it provides distinct computational tools for informing cybersecurity decisions [Kamhoua 2021]. The chief benefit of applying MARL to cybersecurity is that it enables *adaptive* decision making. As the problem and adversary change, MARL provides evolving strategies. MARL may also be used to find *good* strategies when the complexity of the problem makes it impractical to find optimal ones.

4.3.1 Academic Research on Building APT Defenses

The bibliometric analysis reported in Chapter 2 found 49 articles on game theory defenses for APTs and 14 articles on MARL defenses.[7] Following the application-based classification scheme that Kumar and their coauthors use [Kumar 2021], we placed each article into one of the following four categories:

- **Resource allocation.** Distributing resources to reduce the likelihood of successful attacks
- **Deception.** Using honeypots, moving target defenses, and other forms of deception to propagate false beliefs and cause attackers to act against their vested interests
- **Information leakage.** Implementing defenses to reduce the inadvertent disclosure of sensitive information to an attacker
- **Optimal design.** Balancing the security benefits of reactive measures, (e.g., scanning, backing up, or restoring nodes) with financial costs and a system's overall quality of performance

Figure 7 shows the number of game theory and MARL papers in each category. There were far more game theory papers; this was due, in part, to the fact that game theory has a long history, whereas MARL only emerged recently. Both game theory and MARL papers overwhelmingly focused on optimal design and resource allocation. Finally, articles on resource allocation and deception tended to focus on preventing intrusions, whereas articles on information leakage and optimal design tended to focus on mitigating damage from intrusions. Thus, a near equal share of papers concerned proactive and reactive APT defenses.

[7] Seven articles included both game theory and MARL. This is because MARL can be applied to many of the same problems as game theory and can give approximate solutions when it is infeasible to find optimal ones using game theory.

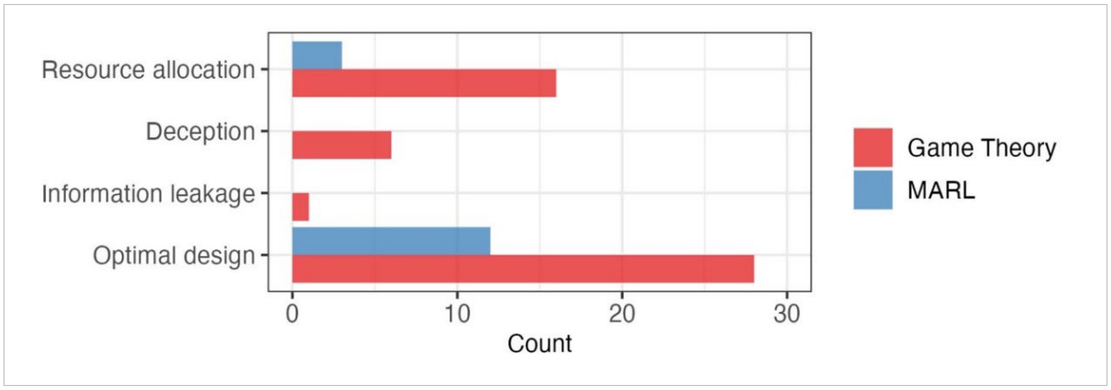

Figure 7: Number of Academic Papers About Game Theory and MARL Defenses Against APTs

Many canonical problems studied in game theory make the following strong assumptions:

- The game includes exactly two players.
- Players have complete and perfect information about game objectives, actions, and payoffs.
- The game is unchanging.
- Players are rational.

These assumptions rarely hold for defending against APTs. Thus, these assumptions must be relaxed to create valid APT games. The articles we reviewed addressed these issues to varying degrees.

Multiple Attackers

An organization must defend against multiple attackers, each with different objectives and strategies. This complicates matters due to the exponential growth in the size of the joint action space the defender must consider. Ultimately, no single defense strategy can be effective against all attackers.

- **Game theory** researchers have analyzed multi-player games, but the resulting methods have not been widely applied in the cyber domain [Papadimitriou 2005]. However, some work on multi-player cyber defense problems has explored how multiple defenders can exchange information to derive strategies from their collective experience [Zhu 2022].
- **MARL** does not depend on a model of the attacker and can be applied to problems with multiple attackers [Zhu 2022]. However, as the number of attackers increases, more training might be needed to arrive at a high-performing solution.

Uncertainty

Due to the stealthy nature of APTs, a defender might have only partial information about compromises already present in a system. Further, a defender might have only partial information about an attacker's identity, motivation, and capabilities. In other words, the defender must make decisions with *imperfect* and *incomplete* information.[8]

- **Game theory** literature related to APTs contains examples of games with imperfect and incomplete information. Many of these papers solve for a *Bayesian Nash* equilibrium, a strategy profile that maximizes expected utility subject to a player's beliefs about the state of the environment and about other players, which may be uncertain [Huang 2020, Kinneer 2019]. Other papers address methods for gathering information to reduce uncertainty about the attacker [Kinneer 2019]. Importantly, uncertainty in APT games almost always relates to the attacker type. Real-world problems include additional sources of uncertainty, such as sensor noise, behavior of third-party players, and other chance events. Accounting for these forms of uncertainty yields better quality solutions, but it may be computationally intractable.
- **MARL** does not depend on a model of the attacker or the environment. Given enough training, it will arrive at a high-performing solution without resolving these sources of uncertainty. However, if the attacker or the environment changes, the model must be (partially) retrained.

Multi-Stage Games

APT attacks are, by definition, multi-stage. As the attacker progresses toward an objective, the state of the problem changes. This in turn necessitates dynamic defense strategies.

- **Game theory** papers related to APTs represent games as a branching tree of choice points, which they solve by decomposing outcomes into immediate and future-stage rewards [Huang 2018, Xiao 2020]. However, while exact methods exist, they have high computational complexity and may not be scalable for real-world APT defenses.
- **MARL** learns the values of actions in particular states, which naturally correspond to different stages in a game. The values reflect the immediate and discounted future reward expected following an action. Several APT papers use learning-based approaches to find good policies for multi-stage games [Moothedath 2020, Sahabandu 2020]. However, these policies are not provably optimal.

Repeated Play

Because APTs are persistent, an attacker may repeatedly interact with a defender. The attacker may adopt new strategies across the sequence of interactions. Likewise, the defender may adopt new strategies based on information learned about the attacker.

- **Game theory** enables the development of games that involve repeated choices. One approach used in these games is to gather information during each interaction to update expectations

[8] In games with incomplete information, players do not have full information about their opponents or about the structure of the game itself. In games of imperfect information, some aspects of play, such as an opponent's previous moves, are hidden from view.

about the attacker type [Bakker 2020, Halabi 2021]. In this way, the defender can adopt tailored strategies.

- **MARL** enables the development of repeated-choice games that update the estimated utility of taking different actions based on experience [Abass 2017, Moothedath 2020, Sahabandu 2020]. Asymptotically, these learning-based approaches arrive at optimal strategies. Further, in non-stationary environments, they continue to shift to new strategies as the environment changes. A drawback is that they are sample inefficient, so they may require a tremendous amount of training to approach an acceptable level of performance.

Subjective Utility

As human beings or nation states, APT attackers are not rational in the sense of maximizing expected utility. For example, an attacker may select actions that yield immediate rewards at the expense of larger future rewards. Deviations from expected utility theory are well documented in the behavioral sciences [Kahneman 2013], and these findings have implications for APT defenses.

- **Game theory** literature related to APTs includes two forms of deviation from rationality. The first relates to suboptimal decisions arising from players' mistaken beliefs about unobservable information, such as the type of opponent they are facing. Methods for dealing with imperfect and incomplete information permit defenders (and attackers) to make *boundedly rational* decisions in these cases. The second form of deviation relates to subjective evaluations of variables like value, probability, and time. With respect to value and probability, several papers use prospect theory to model an asymmetric function in which losses "feel" worse than equivalently sized gains [Tian 2020, Xiao 2018, Xu 2016]. With respect to time, some papers have applied exponential discounting to future rewards to assign greater weight to near-term outcomes [Merlevede 2021, Van Dijk 2012].
- **MARL** does not depend on a model of the attacker. Given enough training, it may arrive at a high-performing solution that exploits an attacker's cognitive biases.

4.3.2 Future Research Directions for Game Theory and MARL to Aid APT Defense

Aspects of APTs can be framed as games. By doing so, mathematical and computational tools from game theory and MARL can be applied to APT defense. Notwithstanding the significant progress in this area, several barriers must be overcome to transition game theory and MARL research into practice. Table 5 summarizes challenges for active APT defenses along with potential solutions.

Table 5: Challenges and Solutions for Active APT Defenses

Challenge	Description	Solutions
Acquiring game models	Game theory methods depend on game definitions, including actors, actions, states, rewards, and preferences. Although MARL does not directly operate on these definitions, a game model is still needed to create the simulation environment and train the agent. In cyber defense, the decision and action space may be especially large and hard to define. Further, it may be difficult to assign objective values to certain outcomes in the cyber domain.	• Maintain system-design documents to characterize the environment. • Conduct workshops with diverse stakeholders to define the benefits and consequences (i.e., utility) of different outcomes.
Scalability	Given the complexity of cyber defense problems and the time-sensitive nature of responses, game theory methods must employ suitable simplifications to arrive at acceptable quality solutions in a reasonable amount of time. MARL solutions leverage pre-trained networks, and can thus act quickly at the time of deployment. Still, as the complexity of the problem increases, it may become difficult to train MARL systems.	• Evaluate the time complexity of methods and the feasibility of deploying them on different computational resources. • Adopt methods that constitute an attractive tradeoff between speed and accuracy.
Explainability	Deep neural networks are black-box models, meaning that humans, even those who design them, may not understand how they arrive at their outputs. All the MARL papers we reviewed used deep reinforcement learning. Thus, the logic underlying MARL policies are hard to explain to human decision makers, let alone to formally verify.	• Develop and integrate explainability interfaces with deep learning methods.
Benchmarking	Of the over 60 game theory and MARL papers we evaluated, one used a fielded setting to validate its solution, 11 used medium-fidelity simulation environments, 30 used low-fidelity simulation environments, and the rest presented theoretical forms of validation. The lack of standardized environments, metrics, and tools for evaluating game theory and MARL solutions for APTs makes it difficult to determine their relative effectiveness and technological maturity.	• Reach consensus in the cyber-defense community about relevant MoPs and MoEs. • Evaluate methods using agreed-on problems and simulation environments.

4.4 Summary

Academic research on AI-enabled APT defense addresses three challenges:

1. detecting threats
2. making sense of multiple threat indicators
3. selecting effective and efficient defense policies

In the academic literature we reviewed, most methods for detection and sensemaking use supervised and unsupervised learning, whereas most methods for selecting policies use game theory or MARL. The performance of AI in academic papers is promising, yet several challenges must be overcome to transition these algorithms to practice.

5 Practical Steps Toward Using ML for APT Defense in Real Networks

Based on our landscape analysis and literature review, we recommend that organizations consider leveraging AI for APT defense for the following reasons:

- Many of the papers we reviewed include computational experiments on synthetic or real data. The performance of AI methods reported in these papers is promising, suggesting that AI may ultimately play an important defense-in-depth role in APT defense.
- Certain APT attributes require more flexible approaches to detection, sensemaking, and planning than traditional methods offer. For example, zero-day exploits, which differ from past attacks, may evade detection by signature-based methods. However, anomaly-based detection methods using AI are well established in many fields and are designed to detect deviations from normalcy without relying on past event signatures [Chandola 2009]. As such, these methods may also be useful for detecting anomalies resulting from novel attacks.
- Building AI solutions for a specific network has several indirect benefits, including increasing the organization's knowledge of its cyber defenses and vulnerabilities. These benefits can be realized even if the AI solutions are not fully deployed.

In this chapter, we present a general framework for incorporating AI into APT defense. Organizations can incrementally apply some, but not all, of the steps contained in the framework. These efforts do not need to be costly, and they can be led by internal cybersecurity personnel with experience in data science.

5.1 Framework for Including Tailored AI Applications into APT Defense

The framework presented in Figure 8 outlines a sequence of steps that organizations can use to adopt AI-enabled approaches in their APT defense. The process starts with defining network-specific problems and ends with deriving effective defense measures and countermeasures. This framework is built on three key ideas:

1. Given the significant differences among organizations, there is no one-size-fits-all AI solution. To develop tailored applications, an organization must first arrive at a network-specific problem formulation.
2. Network-specific problem formulation is the foundation of all subsequent design decisions, and each decision creates the context for downstream decisions.
3. The framework splits into two branches; different decisions lead to the adoption of (1) AI for detection and sensemaking or (2) AI for mitigation.

We describe each of these steps in Sections 5.1.1–5.1.4.

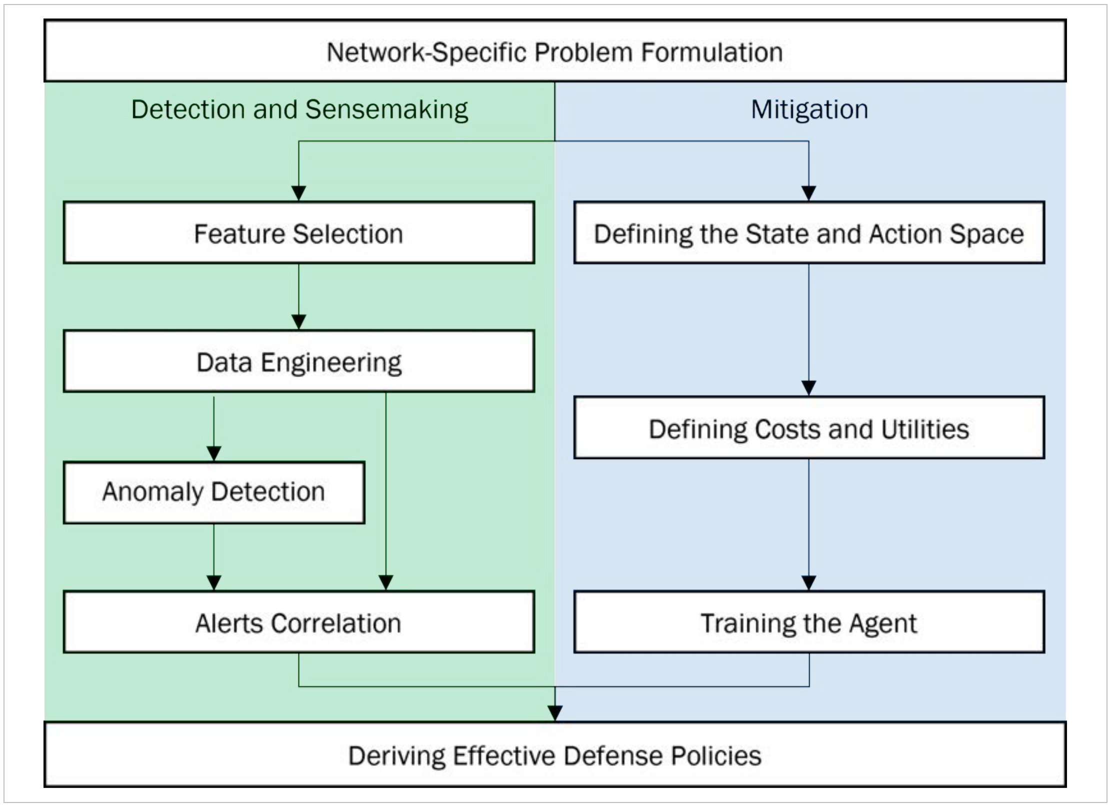
Network-Specific Problem Formulation
Detection and Sensemaking
Mitigation
Feature Selection
Defining the State and Action Space
Data Engineering
Defining Costs and Utilities
Anomaly Detection
Alerts Correlation
Training the Agent
Deriving Effective Defense Policies

Figure 8: Framework for Incorporating AI-Enabled Techniques into APT Defense

5.1.1 Network-Specific Problem Formulation

In the case of detection and sensemaking, the goals of problem formulation are to describe the types of attacks an organization may face as well as the indications and warnings of those attacks. This information might be developed by systematically reviewing the organization's highest value assets (the most likely APT targets) and mapping the steps an attacker might make to reach those assets. This review should consider known vulnerabilities and the weakest links in the network's security. Examples of detectable artifacts include those used by MLAPT: disguised executable file detection, malicious file hash detection, malicious domain name detection, malicious IP address detection, malicious SSL certificate detection, domain flux detection, scan detection, and Tor connection detection [Ghafir 2018].

In the case of mitigation, the goals of problem formulation are to describe the parties who interact with and rely on the network, the timescale of interest, and the timesteps for acting. The timescale of interest (e.g., days, weeks, months) informs assumptions that can be made about whether the environment is stationary or dynamic and whether it can be modeled as a one-shot game or a repeated-play game. It also has implications for the scalability of certain algorithms. The timestep for acting (e.g., continuously, at discrete intervals) has implications for the suitability of different types of algorithms.[9]

5.1.2 Detection and Sensemaking

The detection and sensemaking branch shown in Figure 8 contains four stages:

1. **Feature Selection**. The organization identifies the observable sensor and log file data that can serve as indicators of detectable APT artifacts. These indicators might include output from existing cybersecurity systems. Some APT artifacts might be observable directly, and others may require code to preprocess data into desired formats (e.g., extracting specific information from log files, transforming text to numeric data).

2. **Data Engineering**. The organization develops a system to query the raw data, implements any automated preprocessing needed, and stores the resulting feature data in some type of structured repository. At this exploratory stage, the system can be ad hoc and does not need to operate in real time.

   Given the essential role of data in training ML models, further activities at this stage may involve defining a plan to protect data against adversarial influence [Kurakin 2016].

3. **Anomaly Detection**. ML-based anomaly detection methods have realized value in practical industrial settings outside of cybersecurity. For example, the nuclear power industry has deployed these methods for online condition monitoring of industrial systems and components [Bickford 2002, Davis 2002, Hines 2005]. Given a set of normal operating data, the organization trains an ML model (or models) to predict some response that is monitored by sensors. Once deployed, the trained model makes real-time predictions of the expected response. Large deviations between the predicted response and the true response, as observed by the sensor(s), are flagged as anomalies that require investigation.

   One reason why ML-based anomaly detection is feasible in industrial settings is that the models can focus on small, well-defined systems that have a manageable number of features known to be important. Of course, defining these features is a challenge in large, potentially high-traffic computer networks. However, feasibility improves when the scope of the cyber anomaly detection system can be made smaller and more well defined, especially by leveraging analysts who are experienced in the specific system to be monitored.

4. **Alerts Correlation**. The organization searches for related alerts over time; it is one of the more practical APT detection strategies proposed in the academic research, and it does not necessarily require AI. An appealing aspect of this strategy is that it surrounds an APT's hallmark attribute: low-and-slow correlated events that occur over a long period of time. Furthermore, analysts are unlikely to recognize these sequences of searches without the aid of automated correlation analysis.

[9] Problem formulation may include additional details, such as whether the attacker can first observe defender actions (e.g., via reconnaissance), or whether attackers and defenders must select strategies without first observing one another's decisions.

The process generally involves mapping low-level alerts to the potential stages of an APT attack model, such as the Lockheed Martin Cyber Kill Chain. Then, sequences of low-level events occurring in the proper time order can be identified as potential evidence of an APT that requires investigation. HOLMES [Milajerdi 2019] and MLAPT [Ghafir 2018] are two practical correlation-based APT detection frameworks that can be explored.

5.1.3 Mitigation

The mitigation branch shown in Figure 8 also includes three stages:

1. **Defining the State and Action Space**. The network attack surface is large and, given the multiple configurations possible at each node, high dimensional. Additionally, given the emergence of the Internet of Things, the attack surface may be ill-defined and everchanging.

   To determine its most effective defense policies, the organization must first define the extent of the network to be defended. The organization must then select suitable approximations to represent the state space (i.e., the set of environment conditions that determine the actions to be taken). This selection constitutes a tradeoff. Adopting a more detailed representation may allow highly tailored actions, yet it comes at the expense of increased computational complexity. Adopting a less detailed representation may allow algorithms to find solutions in a feasible amount of time, yet it may yield actions that are not fully optimized to the current environment state.

   Aside from defining the state space, the organization must define the action space. Defensive strategies may include allocating limited resources across nodes, manually inspecting a subset of all indicators, periodically resetting passwords and restoring systems, establishing moving target defenses, partitioning off compromised parts from the rest of the system, reinitializing the system, and deploying honeypots. The levels of each of these strategies determine the action space. The space, though potentially large, is generally known to the defender.[10]

   Offensive strategies include different attack vectors. The levels of each of these strategies and across the set of attackers may be large and are only partially known to the defender. This uncertainty complicates applying game theory, but it does not preclude using MARL.

2. **Defining Costs and Utilities**. In one set of cyber defense problems, the defender must allocate a fixed pool of resources in a near-optimal manner. To approach these problems, the organization must establish the significance of different nodes in the network and the consequences of failing to defend those nodes. This requires assigning subjective values to qualitatively different outcomes (e.g., disruption versus data exfiltration). This also requires assigning subjective values to related outcomes (e.g., different types of data that may be exfiltrated).

[10] A concrete example is Libratus, a state-of-the-art game theory agent for playing poker. It uses abstraction to reduce the 2.4 billion hands possible in the fourth round of Texas hold’em to 1.25 million buckets [Brown 2017].

In another set of cyber defense problems, the defender must balance reducing cyber risks with the costs of taking different actions. In addition to establishing the consequences of different cyber breaches, this balancing requires the defender to establish the costs of different actions. These costs may be monetary, relate to the overutilization of the cyber workforce due to inspecting more indicators, or relate to the degraded quality of service due to system downtime and adopting more restrictive measures. A challenge is that the costs of actions—like the consequences of outcomes—are subjective. Further, they are on different scales. Not having a common currency makes it difficult to balance cost and risk.

The attacker must also allocate limited resources to achieve subjectively valued objectives. These costs and rewards are only partially known to the defender. This uncertainty complicates applying game theory, but it does not preclude using MARL.

3. **Training the Agent**. Once the problem formulation is complete, analytic methods can be used to find equilibrium strategies. When using MARL, this can entail training the agent in a simulation environment. Hyperparameters (i.e., training parameters that affect the quality and speed of learning) are crucial in deep reinforcement learning. Thus, problem formulation and hyperparameter tuning are *both* essential to training an effective agent.

5.1.4 Deriving Effective Defense Policies

The branches in Figure 8 converge on deriving effective defense policies. These policies may come in one of three forms:

- **Detection and Sensemaking only (left path).** An organization may use AI to create systems for detecting threats and for integrating indicators to arrive at a holistic assessment of threat activity.
- **Mitigation only (right path).** An organization may use AI to select a defensive strategy without using AI-enabled techniques for threat detection.
- **Integrated Policy (both paths).** An organization may apply AI for detection, sensemaking, and acting. These two sets of activities can occur in parallel. Alternatively, outputs from the detection and sensemaking paths (i.e., indicators and warnings) can be passed to the state-space representation in the action path. In the latter, AI can be used to decide whether and how to respond to different indicators.

5.1.5 Summary

Commercial development and academic research lay the foundation for two classes of opportunities to integrate AI into APT defense. The first class applies supervised and unsupervised learning to threat detection and sensemaking. Using AI for anomaly detection (i.e., detection) is well established and provides near-term opportunities for organizations. AI has not yet been widely used for event correlation (i.e., sensemaking). Thus, its use reflects a mid-term opportunity. However, given that there are fewer existing methods for sensemaking, applying AI to this problem may be more transformative than applying it to anomaly detection.

The second class of opportunities deals with establishing proactive and reactive defenses. Once again, AI has not yet been widely used for action selection in cyber defense. Given the consequential nature of defender actions, it is unlikely that AI would be used to fully automate decision making soon. However, using AI for human augmentation in APT decision making reflects a mid- to far-term opportunity.

6 Conclusion

The digital landscape of the 21st century has been marked by a significant increase in cyber threats, particularly APTs. These sophisticated, stealthy, and continuous attacks pose a grave risk to organizations, often evading detection by traditional cyber-defense techniques. However, the rise of AI presents a promising solution to this growing concern. AI, with its ability to learn and adapt, offers a robust defense-in-depth strategy against APTs. Yet the transition from research to practice is not without its challenges.

To effectively leverage AI in cyber defense, we must first address several key issues. These include achieving acceptable levels of accuracy for high-stakes cyber defense problems, reducing time complexity to enable real-time decision making, scaling to larger problems, and increasing interpretability. These challenges are not insurmountable, but they require concerted effort and collaboration within the research community.

From a research perspective, four recommendations emerge:

1. **Agree on a set of performance metrics.** The common metric used to assess AI systems—accuracy—is necessary but not sufficient for deciding whether to field a system. In the case of APTs, other relevant metrics include time complexity, memory demands, training time, and robustness to changing inputs.
2. **Evaluate algorithms using more representative data sets and test cases.** Much of the work on AI systems for APT arises from academic research. Evaluating the resulting systems with test cases that reflect the complexity inherent in real-world applications, including uncertainty, noisy inputs, and a dynamic environment, would advance their technology readiness levels. On the one hand, benchmark data sets and test cases should be general enough to evaluate methods arising from diverse sources. On the other hand, the more tailored they are to a specific organization, the stronger the performance guarantees they permit. That said, given the evolving nature of APTs, results obtained with representative data and test cases limit AI system performance.

3. **Continue to explore applying AI techniques from other domains to cyber defense.** Existing work on supervised and unsupervised learning has proven fruitful for threat detection, and game theory and MARL have shown promise for defense planning. A natural next step would be to demonstrate that established academic results can be replicated in operational settings. The fusion of threat indicators remains relatively unexplored. Bayesian methods and cognitive psychology literature on sensemaking may offer valuable insights into automating this process.

4. **Develop methods to increase the explainability of AI systems.** The performance of an AI system, though paramount, is not the only consideration for its adoption. Because AI systems operate alongside human analysts, explainability is also an important consideration. Methods for increasing explainability must be developed and integrated with AI-enabled APT defenses.

From an organizational perspective, the adoption of AI into defense strategies against APTs can begin as research advances. AI methods can address significant gaps in existing defenses, and the process of tailoring these defenses to an organization can provide valuable insights into its cyber defense posture.

Two additional recommendations emerge for organizations:

1. **Prioritize adopting AI for threat and anomaly detection.** These techniques are mature and have been successfully demonstrated in other applied contexts. Additionally, sensemaking requires reliable indications and warnings. Finally, since adopting AI for this application does not directly involve implementing countermeasures as in the case of AI systems for action selection, the risks are lower. Some of the primary considerations while adopting AI systems for threat and anomaly detection are to tailor them to specific organizations, seek an acceptable level of false alarms, and evaluate them in representative conditions.

2. **Integrate AI into a layered strategy that includes conventional threat detection methods and human analysts.** AI methods may address a gap, but they are fallible. Thus, they are best seen as an element of a defense-in-depth strategy. AI methods adopted for threat and anomaly detection can be evaluated in isolation. However, the integrated collection of defense strategies should be holistically evaluated as well.

Finally, we briefly consider the APT vulnerabilities and defense opportunities presented by generative AI. Today, without any technical expertise in AI and ML, threat actors can use open source tools to create deepfake videos, audio, and text that encourages targeted people (e.g., employees in an organization) to reveal information or take some action. For example, in 2022 a deepfake video of Ukrainian President Volodymyr Zelenskyy directed his country’s soldiers to surrender to Russian forces [Allyn 2022]. In 2024, a robocall containing an AI-generated impersonation of United States President Joe Biden circulated in an apparent attempt to suppress voting in New Hampshire [Swenson 2024]. Also in 2024, deepfake images of United States presidential candidate Donald Trump surrounded by Black voters circulated in an apparent attempt to encourage African Americans to vote Republican [Spring 2024]. LLMs can also be used to generate convincing fake news articles and effective clickbait [Chen 2023].

Defending against APTs that leverage generative AI should currently focus on forged media detection. Methods used to detect media created by generative AI is a significant area of active research, and some detection tools are currently available. For example, in July 2024, the Semantic Forensics program of the Defense Advanced Research Projects Agency published a catalog that includes methods to identify AI-generated text, images, audio, and video [DARPA 2024]. While more research is needed to improve reliability, organizations that might be vulnerable should consider using (or at least tracking) available detection methods.

Opportunities for using generative AI to defend against APTs, regardless of whether the threat actor uses generative AI, could be developed with additional research. For example, given a sequence of words, LLMs can predict the most likely sequence of words to follow; for example, this capability can be used to auto-complete text in word processors and computer code in integrated development environments. With additional research, language-based models might be trained to recognize normal sequences of logs for network traffic and/or endpoint activity. Such models might be used as anomaly detectors to identify unusual log sequences. Additional areas where generative AI might assist APT defense include automating the creation of convincing honeypots, having LLMs analyze security-related code, and assisting security teams with triaging alerts and monitoring through the summarization capabilities of LLMs.

In conclusion, the rise of APTs presents a significant challenge to organizations, but the growth of AI offers a promising solution. By addressing key challenges and following the recommendations outlined above, organizations and the research community can work together to harness the potential of AI in cyber defense.

Appendix

Game Definitions

A game consists of three elements:

1. the set of players
2. the set of actions each player may take
3. the payoff function that describes the payoffs each player receives based on the state of the game and the actions chosen by other players

Table 6 shows the Prisoner’s Dilemma, a classic two-player zero-sum game. The two players are denoted by the rows and columns. Each player may choose between two actions: cooperate or defect. The payoff function is given by the cell values. If both players cooperate, each suffers a small loss (-1). If both players defect, each suffers a moderate loss (-2). However, if one player cooperates and the other defects, the player who cooperates suffers a large loss (-3), while the other loses nothing.

Defection is the strictly dominant strategy; it always results in a better outcome than if the player had cooperated. Thus, mutual defection is the Nash Equilibrium. It is the one outcome from which a player could only do worse by unilaterally changing their decision. Based on this analysis, a rational player would defect.

Table 6: Prisoner’s Dilemma

	Cooperate	**Defect**
Cooperate	(-1, -1)	(-3, 0)
Defect	(0, -3)	(-2, -2)

In the context of cybersecurity, the three elements of games can be operationalized as follows:

- **Players.** At a minimum, players include defenders and one or more attackers.
- **Actions.** Defenders must decide which nodes or devices to protect or scan, how to use deceptive tactics like honeypots, and how and when to deploy moving-target defenses. Attackers must decide which nodes or devices to attack and whether and how to exfiltrate data.
- **Payoffs.** The defender’s payoffs relate to the consequences of failing to prevent a cyber attack. Payoffs may also reflect costs (i.e., resources or degraded quality of service) associated with different defense strategies. The attacker’s payoffs relate to the benefits of conducting a successful attack along with the consequences of being detected. Payoffs may also reflect the resources needed for different types of attacks.

MARL problems include, but are not limited to, games like the Prisoner’s Dilemma shown in Table 6. As with games, the specification of a MARL problem includes players, actions, and outcomes. MARL problems also typically specify the set of world states. Not only do joint actions produce rewards or penalties, but they may also change the context for future decisions. Lastly, MARL problems typically specify probabilities to capture uncertainty in the outcomes produced by different joint actions.

The key difference between game theory and MARL is in how solutions are reached. In game theory, solutions are derived from mathematical analysis of the problem specification. In contrast, MARL solutions are derived through trial-and-error interactions between agents in a simulation environment based on the problem specification.

Glossary

Advanced Persistent Threat (APT) Attack

An attack carried out by an APT group

Advanced Persistent Threat (APT) Group

An adversary with sophisticated levels of expertise and significant resources, allowing it—through the use of multiple different attack vectors (e.g., cyber, physical, and deception)—to generate opportunities to achieve its objectives, which are typically to establish and extend footholds within the information technology infrastructure of organizations for purposes of continually exfiltrating information and/or to undermine or impede critical aspects of a mission, program, or organization or place itself in a position to do so in the future; moreover, pursuing the advanced persistent threat's objectives repeatedly over an extended period of time, adapting to a defender's efforts to resist it, and determining to maintain the level of interaction needed to execute its objectives

Anomaly Detection

A subfield of ML concerned with producing models that identify abnormal inputs

Artificial Intelligence (AI)

A subfield of computer science concerned with designing and building intelligent agents that receive percepts from the environment and take actions that affect that environment

Attack Campaign

An ongoing series of cyber attacks or attempted cyber attacks committed by an actor against one or more targets

Backdoor

An undocumented way of gaining access to a computer system

Benchmark

A fixed set of environmental factors and quality metrics used to make objective comparisons

Classifier

A model that associates inputs with output categories or a probability distribution over those categories

Clustering

An ML algorithm that takes inputs and finds a natural grouping of the input data into separate clusters

Computer Vision

A subfield of AI that uses computer programs to perform tasks of visual cognition

Convolutional Neural Network

A neural network with convolutional layers that apply a transformation at each location in the input

Cyber Kill Chain

A model for the sequence of actions an attacker must carry out over the course of a cyber attack

Cybersecurity

Prevention of damage to, protection of, and restoration of computers, electronic communications systems, electronic communications services, wire communication, and electronic communication, including information contained therein, to ensure its availability, integrity, authentication, confidentiality, and nonrepudiation

Deep Learning

A type of ML that uses neural networks with multiple hidden layers

Endpoint

A remote computing device that communicates back and forth with a network to which it is connected

Ensemble

A model that consists of a combination of models and a rule for how to combine the results of those models

Exploit

A piece of software, data, or a sequence of commands that takes advantage of a bug or vulnerability to cause unintended or unanticipated behavior to occur on systems that use or are enabled by cyber resources

Feature

A property of the raw data present in the observations that comprise a data set

Feed-Forward Neural Network

A neural network that sequentially applies each layer to the output of the previous layer

Game Theory

A subfield of mathematics that studies the problem of decision making under uncertainty with emphasis on the interactions between decision-making beings (i.e., agents)

Generative AI

Subset of AI that deals with models that can generate data or content that appears authentic or human generated

Graph

A mathematical structure that is widely used in computer science (A graph consists of a set of vertices and a set of edges, where each edge consists of an ordered pair of vertices. Edges may be directed or undirected. If edges are directed, the first vertex in the edge points to the second [e.g., the edge (a, b) means "a points to b"]. If edges are undirected, the first vertex and the second are connected, [e.g., both edges (a, b) and (b, a) mean "a and b are connected"].)

Honeypot

A decoy system or server deployed alongside production systems within a network to act as an enticing target for attackers

Hyperparameter

A parameter that determines the learning process of an AI or ML algorithm

Intrusion Detection System (IDS)

A security technology used to detect and respond to unauthorized activities and potential security threats

Log File

A computer-generated data file that contains information about usage patterns, activities, and operations within an operating system, application, server, or other device

Machine Learning (ML)

An interdisciplinary field combining computer science, mathematics, and statistics to produce generalizations, representations, and predictions from data

Malware

Hardware, firmware, or software that is intentionally included or inserted in a system for a harmful purpose

Markov Decision Problem

A problem from optimal control theory that underlies much of reinforcement learning and AI (The problem is the determination of an optimal sequence of actions in a dynamic system with discrete time increments, where the next system state or the random process that determines the next state depends only on the current state and action taken the current state.)

Memory

A device that is used to store data or programs (e.g., sequences of instructions) on a temporary or permanent basis for use in an electronic digital computer (Typically, the term *memory* is used synonymously with random access memory [RAM], a type of computer memory that can be searched in any order and changed as necessary.)

Metric

A numeric measure that represents some aspect of a model

Model

An abstraction or approximation of some aspect of the world (In the context of AI and ML, models are used to produce predictions or actionable information.)

Multi-Agent Reinforcement Learning

A reinforcement learning paradigm in which multiple agents are cooperating or competing with one another

Natural Language Processing

A subfield of AI that uses computer programs to perform intelligent tasks using text expressed in a natural language (e.g., English, Spanish)

Network

Information system(s) implemented with a collection of interconnected components, which may include routers, hubs, cabling, telecommunications controllers, key distribution centers, and technical control devices

Network Node

An individual device or system within a network

Network Traffic Flow

A sequence of IP packets with protocol headers being added to the payload data to facilitate routing and delivery through the network

Neural Network

A type of ML model based on a mathematical abstraction of the human brain, consisting of multiple layers with interconnections between them

Packet

The basic unit of communication over a digital network

Payload

The portion of transmitted data that is used to fulfill the purpose of the transmission

Phishing

Tricking individuals into disclosing sensitive personal information through deceptive computer-based means

Predictor

A feature or combination of features that have been transformed for use in building an AI or ML model

Q-Learning

A form of reinforcement learning in which the agent learns a Q-table that approximates the cumulative reward expected for taking different actions in each state

Recurrent Neural Network

A neural network that allows connections from later layers to earlier layers or a single layer to itself

Regressor

A model that associates inputs with a numeric output or a probability distribution over numeric outputs

Reinforcement Learning

An area of ML that uses ML algorithms to approximate solutions to optimal control problems

Source Code

A text listing of commands to be compiled, assembled, or interpreted into an executable computer program

Tactics, Techniques, and Procedures (TTPs)

The behavior of an actor (A tactic is the highest level description of this behavior, while techniques give a more detailed description of behavior in the context of a tactic and procedures give an even lower level, highly detailed description in the context of a technique.)

Threat Intelligence

Threat information that has been aggregated, transformed, analyzed, interpreted, or enriched to provide the necessary context for decision-making processes

Vulnerability

Weakness in an information system, system security procedure, internal control, or implementation that could be exploited or triggered by a threat source

Zero-Day Attack

An attack that exploits a previously unknown hardware, firmware, or software vulnerability

Zero-Day Exploit

An exploit that utilizes a zero-day vulnerability

Zero-Day Vulnerability

A previously unknown hardware, firmware, or software vulnerability

Bibliography

URLs are valid as of the publication date of this report.

[Abass 2017]

Abass, Ahmed A. Alabdel; Xiao, Liang; Mandayam, Narayan. B.; & Gajic, Zoran. Evolutionary Game Theoretic Analysis of Advanced Persistent Threats Against Cloud Storage. *IEEE Access*. Volume 5. April 5, 2017. Pages 8482–8491. https://doi.org/10.1109/ACCESS.2017.2691326

[Aburomman 2017]

Aburomman, Abdulla Amin & Reaz, Mamun Bin Ibne. A Survey of Intrusion Detection Systems Based on Ensemble and Hybrid Classifiers. *Computers & Security*. Volume 65. March 2017. Pages 135–152. https://doi.org/10.1016/j.cose.2016.11.004

[Allyn 2022]

Allyn, Bobby. *Deepfake Video of Zelenskyy Could Be 'Tip of the Iceberg' in Info War, Experts Warn.* NPR Website. March 16, 2022. https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia

[Alshamrani 2019]

Alshamrani, Adel; Myneni, Sowmya; Chowdhary, Ankur; & Huang, Dijang. A Survey on Advanced Persistent Threats: Techniques, Solutions, Challenges, and Research Opportunities. *IEEE Communications Surveys & Tutorials*. Volume 21. Issue 2. January 9, 2019. Pages 1851–1877. https://doi.org/10.1109/COMST.2019.2891891

[Bakker 2020]

Bakker, Craig; Bhattacharya, Arnab; Chatterjee, Samrat; & Vrabie, Draguna L. Hypergames and Cyber-Physical Security for Control Systems. *ACM Transactions on Cyber-Physical Systems*. Volume 4. Issue 4. June 18, 2020. Pages 1–41. https://doi.org/10.1145/3384676

[Bhuyan 2013]

Bhuyan, Monowar H.; Bhattacharyya, D. K.; & Kalita, J. K. Network Anomaly Detection: Methods, Systems and Tools. *IEEE Communications Surveys & Tutorials*. Volume 16. Issue 1. June 6, 2013. Pages 303–336. https://doi.org/10.1109/SURV.2013.052213.00046

[Bickford 2002]

Bickford, Randall; Davis, Eddie; Rusaw, Richard; & Shankar, Ramesh. Development of an Online Predictive Monitoring System for Power Generating Plants. Number 805. In *45th ISA POWID Symposium*. 2002. https://www.academia.edu/75327817/Development_of_an_Online_Predictive_Monitoring_System_for_Power_Generating_Plants?uc-sb-sw=33531929

[Borgman 2005]

Borgman, Christine L. & Furner, Jonathan. Scholarly Communication and Bibliometrics. *Annual Review of Information Science and Technology*. Volume 36. February 1, 2005. Pages 1–53. https://doi.org/10.1002/aris.1440360102

[Brown 2017]

Brown, Noam & Sandholm, Tuomas. Superhuman AI for Heads-Up No-Limit Poker: Libratus Beats Top Professionals. *Science*. Volume 359. Issue 6374. December 17, 2017. Pages 418–424. https://doi.org/10.1126/science.aao1733

[Buczak 2015]

Buczak, Anna L. & Guven, Erhan. A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. *IEEE Communications Surveys & Tutorials*. Volume 18. Issue 2. October 26, 2015. Pages 1153–1176. https://doi.org/10.1109/COMST.2015.2494502

[Chandola 2009]

Chandola, Varun; Banerjee, Arindam; & Kumar, Vipin. Anomaly Detection: A Survey. *ACM Computing Survey*s. Volume 41. Issue 3. July 30, 2009. Pages 1–58. https://doi.org/10.1145/1541880.1541882

[Chen 2023]

Chen, Canyu & Shu, Kai. *Combating Misinformation in the Age of LLMs: Opportunities and Challenges*. Illinois Institute of Technology. November 9, 2023. https://arxiv.org/pdf/2311.05656

[CIS 2023]

Center for Internet Security (CIS). *The Cost of Cyber Defense: CIS Controls IG1*. CIS. August 2023. https://www.cisecurity.org/insights/white-papers/the-cost-of-cyber-defense-cis-controls-ig1

[Clarivate 2024]

Clarivate. Web of Science. *Clarivate Website*. March 12, 2024 [accessed]. https://mjl.clarivate.com/home

[DARPA 2024]

Defense Advanced Research Projects Agency (DARPA). Semantic Forensics Analytic Catalog. *DARPA Semantic Forensics Portal*. July 17, 2024 [accessed]. https://semanticforensics.com/analytic-catalog

[Davis 2002]

Davis, E.; Bickford, R.; Colgan, P.; Nesmith, K.; Rusaw, R.; & Shankar, R. On-Line Monitoring at Nuclear Power Plants-Results from the EPRI On-Line Monitoring Implementation Project. Pages 2–7. In *45th ISA POWID Symposium*. 2002. https://www.academia.edu/75327759/On_Line_Monitoring_at_Nuclear_Power_Plants_Results_From_the_EPRI_On_Line_Monitoring_Implementation_Project

[Ferrag 2020]

Ferrag, Mohamed Amine; Maglaras, Leandros; Moschoyiannis, Sotiris; & Janicke, Helge. Deep Learning for Cyber Security Intrusion Detection: Approaches, Datasets, and Comparative Study. *Journal of Information Security and Applications*. Volume 50. February 2020. Page 102419. https://doi.org/10.1016/j.jisa.2019.102419

[GAO 2021]

United States Government Accountability Office (GAO). SolarWinds Cyberattack Demands Significant Federal and Private-Sector Response. *GAO Website.* April 22, 2021. https://www.gao.gov/blog/solarwinds-cyberattack-demands-significant-federal-and-private-sector-response-infographic

[Ghafir 2018]

Ghafir, Ibrahim; Hammoudeh, Mohammad; Prenosil, Vaclav; Han, Liangxiu; Hegarty, Robert; Rabie, Khaled; & Aparicio-Navarro, Francisco J. Detection of Advanced Persistent Threat Using Machine Learning Correlation Analysis. *Future Generation Computer Systems*. Volume 89. December 2018. Pages 349–359. https://doi.org/10.1016/j.future.2018.06.055

[Google 2023]

Google LLC. MLOps: Continuous Delivery and Automation Pipelines in Machine Learning. *Google Cloud Architecture Center Website*. May 18, 2023. https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning

[Halabi 2021]

Halabi, Talal; Wahab, Omar Abdel; Al Mallah, Ranwa; & Zulkernine, Mohammad. Protecting the Internet of Vehicles Against Advanced Persistent Threats: a Bayesian Stackelberg Game. *IEEE Transactions on Reliability*. Volume 70. Issue 3. January 14, 2021. Pages 970–985. https://doi.org/10.1109/TR.2020.3046688

[Harry 2018]

Harry, Charles & Gallagher, Nancy. Classifying Cyber Events: A Proposed Taxonomy. *Journal of Information Warfare.* Volume 17. Issue 3. July 1, 2018. Pages 17–31. https://www.jinfowar.com/journal/volume-17-issue-3/classifying-cyber-events-proposed-taxonomy

[Hines 2005]

Hines, J. W. & Davis, E. Lessons Learned from the U.S. Nuclear Power Plant On-Line Monitoring Programs. *Progress in Nuclear Energy*. Volume 46. Issues 3-4. August 5, 2005. Pages 176–189. https://doi.org/10.1016/j.pnucene.2005.03.003

[Huang 2018]
Huang, Linan & Zhu, Quanyan. Analysis and Computation of Adaptive Defense Strategies Against Advanced Persistent Threats For Cyber-Physical Systems. Pages 205–226. In *Decision and Game Theory for Security GameSec 2018 Lecture Notes in Computer Science*. Volume 11199. September 2018. https://doi.org/10.1007/978-3-030-01554-1_12

[Huang 2020]
Huang, Linan & Zhu, Quanyan. A Dynamic Games Approach to Proactive Defense Strategies Against Advanced Persistent Threats in Cyber-Physical Systems. *Computers & Security*. Volume 89. February 2020. Page 101660. https://doi.org/10.1016/j.cose.2019.101660

[Hutchins 2011]
Hutchins, E.; Cloppert, M.; & Amin, R. Intelligence-Driven Computer Network Defense Informed by Analysis of Adversary Campaigns and Intrusion Kill Chains. *Leading Issues in Information Warfare & Security Research*. Volume 1. Issue 1. 2011. Page 80. https://community.mis.temple.edu/mis5208sp2016/files/2015/01/iciw2011.pdf

[Johnson 2016]
Johnson, Ariana L. Cybersecurity for Financial Institutions: The Integral Role of Information Sharing in Cyber Attack Mitigation. *North Carolina Banking Institute Journal*. Volume 20. 2016. Page 277. https://scholarship.law.unc.edu/ncbi/vol20/iss1/15/

[Jordan 2015]
Jordan, M. I. & Mitchell, T. M. Machine Learning: Trends, Perspectives, and Prospects. *Science*. Volume 349. Issue 6245. July 17, 2015. Pages 255–260. https://doi.org/10.1126/science.aaa8415

[Kahneman 2013]
Kahneman, Daniel. (2011). *Thinking, Fast and Slow*. Macmillan Publishers. 2013. ISBN 9780374533557. https://us.macmillan.com/books/9780374533557/thinkingfastandslow

[Kamhoua 2021]
Kamhoua, Charles A.; Kiekintveld, Christopher D.; Fang, Fei; & Zhu, Quanyan, eds. *Game Theory and Machine Learning for Cyber Security*. John Wiley & Sons. 2021. ISBN 978-1-119-72392-9. https://www.wiley.com/en-us/Game+Theory+and+Machine+Learning+for+Cyber+Security-p-9781119723929

[Kinneer 2019]
Kinneer, Cody; Wagner, Ryan; Fang, Fei; Le Goues, Claire; & Garlan, David. Modeling Observability in Adaptive Systems to Defend Against Advanced Persistent Threats. Pages 1–11. In *Proceedings of the 17th ACM-IEEE International Conference on Formal Methods and Models for System Design*. October 2019. https://dl.acm.org/doi/abs/10.1145/3359986.3361208

[Kumar 2021]
Kumar, Rajesh; Singh, Siddhant; & Kela, Rohan. Analyzing Advanced Persistent Threats Using Game Theory: A Critical Literature Review. Pages 45–69. In *Critical Infrastructure Protection XV*. March 2021. https://link.springer.com/chapter/10.1007/978-3-030-93511-5_3

[Kurakin 2016]
Kurakin, Alexey; Goodfellow, Ian; & Bengio, Samy. *Adversarial Machine Learning at Scale*. Cornell University. November 4, 2016. https://doi.org/10.48550/arXiv.1611.01236

[Langner 2011]
Langner, Ralph. Stuxnet: Dissecting a Cyberwarfare Weapon. *IEEE Security & Privacy*. Volume 9. Issue 3. May 23, 2011. Pages 49–51. https://doi.org/10.1109/MSP.2011.67

[Lockheed Martin 2024]
Lockheed Martin Corporation. The Cyber Kill Chain. *Lockheed Martin Website.* March 12, 2024 [accessed]. https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html

[Marotta 2017]
Marotta, Angelica; Martinelli, Fabio; Nanni, Stefano; Orlando, Albina; & Yautsiukhin, Artsiom. Cyber-Insurance Survey. *Computer Science Review*. Volume 24. May 2017. Pages 35–61. https://doi.org/10.1016/j.cosrev.2017.01.001

[McKinsey 2022]
McKinsey & Company. The State of AI in 2022—and a Half Decade in Review. *QuantumBlack AI by McKinsey Website.* December 6, 2022. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2022-and-a-half-decade-in-review

[Merlevede 2021]
Merlevede, Jonathan; Johnson, Benjamin; Grossklags, Jens; & Holvoet, Tom. Exponential Discounting in Security Games of Timing. *Journal of Cybersecurity*. Volume 7. Issue 1. March 17, 2021. https://doi.org/10.1093/cybsec/tyaa008

[Milajerdi 2019]
Milajerdi, S. M.; Gjomemo, R.; Eshete, B.; Sekar, R.; & Venkatakrishnan, V. HOLMES: Real-Time APT Detection Through Correlation of Suspicious Information Flows. Pages 1137–1152. In *2019 IEEE Symposium on Security and Privacy (SP)*. September 2019. https://doi.org/10.1109/SP.2019.00026

[MITRE 2024]
The MITRE Corporation. Groups. *The MITRE ATT&CK Website*. March 12, 2024 [accessed]. https://attack.mitre.org/groups/

[Moothedath 2020]

Moothedath, Shana; Sahabandu, Dinuka; Allen, Joey; Clark, Andrew; Bushnell, Linda; Lee, Wenke; & Poovendran, Radha. A Game-Theoretic Approach for Dynamic Information Flow Tracking to Detect Multistage Advanced Persistent Threats. *IEEE Transactions on Automatic Control*. Volume 65. Issue 12. February 24, 2020. Pages 5248–5263. https://doi.org/10.1109/TAC.2020.2976040

[Myneni 2020]

Myneni, Sowmya; Chowdhary, Ankur; Sabur, Abdulakim; Sengupta, Sailik; Agrawal, Garima; Huang, Dijiang; & Kang, Myong. DAPT2020 - Constructing a Benchmark Dataset For Advanced Persistent Threats. Pages 138–163. In *Deployable Machine Learning for Security Defense: First International Workshop, MLHat 2020.* August 2020. https://link.springer.com/chapter/10.1007/978-3-030-59621-7_8

[NIST 2011]

National Institute of Standards and Technology (NIST). *Managing Information Security Risk: Organization, Mission, and Information System View*. NIST SP 800-39. NIST. 2011. https://csrc.nist.gov/pubs/sp/800/39/final

[NIST 2023]

National Institute of Standards and Technology (NIST). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. NIST. January 2023. https://doi.org/10.6028/NIST.AI.100-1

[Papadimitriou 2005]

Papadimitriou, Christos H. & Roughgarden, Tim. Computing Equilibria in Multi-Player Games. Pages 82–91. In *SODA ‘05: Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. January 2005. https://theory.stanford.edu/~tim/papers/sym.pdf

[Radicati 2021]

The Radicati Group, Incorporated. *APT Protection—Market Quadrant 2021.* March 2021. Radicati. https://www.radicati.com/?p=17305

[Russell 2021]

Russell, Stuart & Norvig, Peter. *Artificial Intelligence: A Modern Approach*. Pearson Education, Incorporated. ISBN 9780137505135. 2021. https://www.pearson.com/en-us/subject-catalog/p/artificial-intelligence-a-modern-approach/P200000003500/9780137505135

[Sahabandu 2020]

Sahabandu, Dinuka; Allen, Joey; Moothedath, Shana; Bushnell, Linda; Lee, Wenke; & Poovendran, Radha. Quickest Detection of Advanced Persistent Threats: a Semi-Markov Game Approach. Pages 9-19. In *2020 ACM/IEEE 11th International Conference on Cyber-Physical Systems (ICCPS)*. May 2020. https://doi.org/10.1109/ICCPS48487.2020.00009

[Samek 2019]

Samek, Wojciech; Montavon, Grégoire; Vedaldi, Andrea; Hansen, Lars Kai; & Müller, Klaus-Robert, eds. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer Publishing. 2019. https://link.springer.com/book/10.1007/978-3-030-28954-6

[Scopus 2024]

Scopus. Scopus Preview. *Scopus Website*. March 12, 2024 [accessed]. https://www.scopus.com/home.uri

[Spring 2024]

Spring, Marianna. Trump Supporters Target Black Voters with Faked AI Images. *BBC Website*. March 2024. https://www.bbc.com/news/world-us-canada-68440150

[Statista 2024]

Statista. Revenue from Advanced Persistent Threat (APT) Protection Market Worldwide from 2015 to 2027. *Statista Website*. March 12, 2024 [accessed]. https://www.statista.com/statistics/497945/advanced-persistent-threat-market-worldwide/

[Swenson 2024]

Swenson, Ali & Weissert, Will. AI Robocalls Impersonate President Biden in an Apparent Attempt to Suppress Votes in New Hampshire. *Public Broadcasting Service (PBS) Website*. January 2024. https://www.pbs.org/newshour/politics/ai-robocalls-impersonate-president-biden-in-an-apparent-attempt-to-suppress-votes-in-new-hampshire

[Tian 2020]

Tian, Wen; Ji, Xiaopeng; Liu, Weiwei; Liu, Guangjie; Zhai, Jiangtao; Dai, Yuewei; & Huang, Shuhua. Prospect Theoretic Study of Honeypot Defense Against Advanced Persistent Threats in Power Grid. *IEEE Access*. Volume 8. April 1, 2020. Pages 64075–64085. https://doi.org/10.1109/ACCESS.2020.2984795

[Tsoularis 2002]

Tsoularis, A. & Wallace, J. Analysis of Logistic Growth Models. *Mathematical Biosciences*. Volume 179. Issue 1. July/August 2002. Pages 21–55. https://www.sciencedirect.com/science/article/abs/pii/S0025556402000962

[Van Dijk 2012]

Van Dijk, Marten; Juels, Ari; Oprea, Alina; & Rivest, Ronald L. FLIPIT: The Game of "Stealthy Takeover." *Journal of Cryptology*. Volume 26. October 26, 2012. Pages 655–713. https://link.springer.com/article/10.1007/s00145-012-9134-5

[Xiao 2018]
Xiao, Liang; Xu, Dongjin; Mandayam, Narayan. B.; & Poor, H. Vincent. Attacker-Centric View of a Detection Game Against Advanced Persistent Threats. *IEEE Transactions on Mobile Computing*. Volume 17. Issue 11. March 8, 2018. Pages 2512–2523.
https://doi.org/10.1109/TMC.2018.2814052

[Xiao 2020]
Xiao, Kaiming; Zhu, Cheng; Xie, Junjie; Zhou, Yun; Zhu, Xianqiang; & Zhang, Weiming. Dynamic Defense Against Stealth Malware Propagation in Cyber-Physical Systems: a Game-Theoretical Framework. *Entropy*. Volume 22. Issue 8. August 2020. Page 894.
https://doi.org/10.3390/e22080894

[Xu 2016]
Xu, Dongjin; Li, Yanda; Xiao, Liang; Mandayam, Narayan. B.; & Poor, H. Vincent. Prospect Theoretic Study of Cloud Storage Defense Against Advanced Persistent Threats. Pages 1–6. In *2016 IEEE Global Communications Conference (GLOBECOM).* February 2016.
https://doi.org/10.1109/GLOCOM.2016.7842178

[Zelonis 2020]
Zelonis, John. *Now Tech: Enterprise Detection and Response, Q1 2020.* Forrester Research, Incorporated. February 5, 2020. https://www.forrester.com/report/now-tech-enterprise-detection-and-response-q1-2020/RES158857?ref_search=0_1681393411587

[Zhu 2022]
Zhu, Tianqing; Ye, Dayong; Cheng, Zishuo; Zhou, Wanlei; & Philip, S. Yu. Learning Games for Defending Advanced Persistent Threats in Cyber Systems. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*. Volume 3. Issue 4. October 19, 2022. Pages 2410–2422.
https://doi.org/10.1109/TSMC.2022.3211866

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave Blank)	2. REPORT DATE August 2024	3. REPORT TYPE AND DATES COVERED Final

4. TITLE AND SUBTITLE
Toward the Use of Artificial Intelligence (AI) for Advanced Persistent Threat Detection

5. FUNDING NUMBERS
FA8702-15-D-0002

6. AUTHOR(S)
Matthew Walsh, Clarence Worrell, Thomas Scanlon

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

8. PERFORMING ORGANIZATION REPORT NUMBER
CMU/SEI-2024-TR-001

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
SEI Administrative Agent
AFLCMC/AZS
5 Eglin Street
Hanscom AFB, MA 01731-2100

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
n/a

11. SUPPLEMENTARY NOTES

12A DISTRIBUTION/AVAILABILITY STATEMENT
Unclassified/Unlimited, DTIC, NTIS

12B DISTRIBUTION CODE

13. ABSTRACT (MAXIMUM 200 WORDS)
This report examines the feasibility and usefulness of implementing artificial intelligence (AI) and machine learning (ML) in cyber defense with a particular focus on advanced persistent threats (APTs). In this report, we examine the current state of AI-enabled APT defense. We begin by describing the stages that an APT must go through to succeed. Next, we perform a commercial market analysis of APT defenses. We then perform a bibliometric analysis to map out the academic research landscape on APTs. We highlight the strengths and limitations of research on the use of AI for APT defense. Finally, we offer practical recommendations that will help organizations start incorporating AI into their layered APT defense strategies.

14. SUBJECT TERMS
advanced persistent threats, artificial intelligence, AI, machine learning, ML

15. NUMBER OF PAGES
60

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT UL

NSN 7540-01-280-5500 Standard Form 298 (Rev. 2-89) Prescribed by ANSI Std. Z39-18 298-102