# Autonomous Adversaries: AI-driven Conflicts in Cybersecurity Systems

Nader Shahata
National Institute of Informatics, Center for Strategic Cyber Resilience Research and Development
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan

**Artificial Intelligence (AI) will create new opportunities, while also creating new challenges for those within the cybersecurity profession as the cybersecurity landscape continues to evolve. The purpose of this paper is to provide a conceptual framework of the evolving dynamics between offensive AI agents (Red AI) and defensive AI agents (Blue AI) taking place in the same cyberspace battlefield. The methods utilized by Red AI to compromise digital assets are varied, including but not limited to, network scanning, exploit execution, and adversarial machine learning. Red AI takes advantage of self-learned strategies (generated by AI) in gaining total control over an organization's networks, devices, data, and applications. Blue AI uses predictive analytics, anomaly detection, and autonomous response strategies to identify, block, and adapt to those attacks. The main battlefield that exists between Red AI and Blue AI is in a cyberspace which included the main components of cyberspace (i.e., Networks, Applications and Data Traffic). This research emphasizes the need for continuing to advance the development of defensive AI technologies to counter Red AI initiated attacks, while also providing a foundation for simulating this adversarial situation (Red AIs attack) against the understanding of future cybersecurity incidents. The outcome of this research is to assist with the enhanced development of AI driven cyber defense systems which ultimately provide a higher level of security for our Network Environment.**

*Index Terms*—**Adversarial AI, Cybersecurity Automation, Network Security, Threat Detection, Machine Learning.**

## I. Introduction

AS the field of cybersecurity continues to grow and the importance of cyberspace structures the social cohesion and national security increases, adversarial threats become increasingly complex. The current conventional network protection methods prove to be inadequate, revealing vulnerabilities to automated attacks. Artificial Intelligence (AI) is emerging as a refined technology used by both attackers and defenders.

Moreover, artificial intelligence becomes increasingly integrated into cybersecurity systems [13], and therefore; new challenges emerge as a result. The conflict race between attackers using AI and defenders using AI will only get faster as both sides are able to leverage machine learning to develop new offensive and defensive tools. For example, autonomous malware can now test networks for vulnerabilities at a surprisingly higher rate than a traditional human analyst can identify. Therefore, AI-driven defense mechanisms are required to predict potential malware and deploy corresponding countermeasure(s) will become a necessity as well as a precaution effort in protecting our data [3]. The consequences of this technical shift will extend beyond the current existing network systems. Autonomous adversaries are raising a lot of critical questions regarding how will the future of cyber warfare will be. The reason is, when an AI system can generate novel attack vectors without the need of human involvement, this will make the current traditional frameworks for understanding and dealing up with cybercrime and cyber warfare become surprisingly insufficient.

The shift from human-operated cyber defense systems to autonomous electronic defense systems represents a change in the way we think about how we identify, deploy and defend against different types of cyber-attacks [3]. In the past, identifying a vulnerability, creating an exploit and executing an exploitation took a lot of time and expertise to accomplish. Cyber adversaries today utilize artificial intelligence to perform most [12], if not all, aspects of cyber-attack execution starting with the reconnaissance phase. Cyber adversaries leverage their ability to optimize their attack strategies by enhancing their knowledge through machine learning [7] techniques and reinforcing feedback from previous attempts to intrude into our network systems.

This paper will explore AI in the context of both conflict and as a protective measure. It investigates the interaction between Red AI (the Attacker) and Blue AI (the Defender) in the cybersecurity domain [2]. Red AI comprises offensive agents that conduct attacks autonomously, employing adversarial recognition, exploitation and machine learning techniques [7]. On the other hand, Blue AI consists of defensive agents that detect and identify anomalies, take preventive measures against intrusions, and formulate an AI-based adaptive security policy, all-in real-time matter. To represent the conflict between AI entities, we developed a visual model called the "Cyberspace Battlefield" [2], which represents an environment filled with networks, applications and data traffic flow. In this situational model, Red AI and Blue AI function as continuous feedback loops of system interactions [2]. Red AI initiates attacks by passing through networks looking for vulnerabilities and executing automated attacks, while Blue AI examines signals and formulates intelligent responses to these threats.

Our architecture adopts a metaphorical and technical perspective to this adversarial interaction; highlighting the importance of machine learning, automation and adaptability in the cyber domain. This paper aims to discuss the benefits, drawbacks, and emerging activities of such autonomous systems, while also emphasizing the need for more research into autonomous cyber defense [6] and AI adversarial threats. The complexity of cyber threats is increasing as cyberspace is

Manuscript received December 2, 2025; revised March 6, 2026. Corresponding author: Nader Shahata (email: nader@nii.ac.jp).

determined to be the battleground for the most sophisticated attacks. Consequently, standard security models are not going be sufficient in the face of highly adaptable and automated attacks. As a result, AI is considered a game changer, on both sides of cyber adversary interaction and conflict, as in the form of weapon and defense in this field.

With the fusion of machine learning and offensive cybersecurity, we are seeing a promising emergence of autonomous adversaries. These autonomous adversaries are capable of learning, adapting and executing attacks autonomously, without the need for direct human oversight. To date, most of the research done in the area of AI-driven exploit generation has treated the attacker as an autonomous decision-making threat actor. However very limited research has been conducted to treat the attacker as a fully autonomous entity acting in a realistic network environment. The lack of such research limits our ability to anticipate new techniques and to build defenses that can cope with self-optimizing threats [4]. Through this research, we hope to increase both the understanding of AI-driven conflict and to provide useful mechanisms to the cybersecurity committees.

## II. Background

The Combination of artificial intelligence and cybersecurity fields represents a significant transformation to both of these domains [14]. The current conventional cyber defense depends on static rules-based systems, human analysts monitoring, and signature-based threat detection. As these methods have been effective against known threats, they are actually inadequate for identifying zero-day exploits, emerging attacks, and sophisticated evasion techniques used by automated adversaries. Red AI (often referred to as offensive AI) has the ability to automate and scale cyberattacks. For example, an AI system can independently scan the protection level of a global network [1], identify exploitable targets, prioritize them, and execute exploitations without the need for human interaction. Additionally, advanced AI models can lead to the formation of adaptive malware that can modify itself to evade detection in real time, resulting in a significant evolution from traditional static ones. Another worrying aspect is adversarial machine learning. An attacker can hide their malicious intentions by manipulating the AI model to make errors. For example, a malware sample can be altered in a way that is undetectable by network analysts, triggering a defense mechanism and being recognized as a threat. This undermines the reliability of AI-based defenses and negatively affects the robustness of the AI model and its anti-adversarial training capabilities [4].

To combat these threats, defenders are increasingly required to deploy Blue AI to improve situational awareness, accelerate response times, and scale security practices. Blue AI employs a combination of supervised and unsupervised learning techniques to gain a better understanding of behaviors in online environments, facilitating anomaly detection, correlating threat indicators, and predicting potential attack incidents. Blue AI can be operating in Security Operations Centers (SOC), endpoint detection and response (EDR) systems, or through cloud platforms, enabling defenders to react to incidents in a real time and automatically implement relevant policies. Many existing cybersecurity blue teams are gradually progressing toward autonomous cyber defense [6], where AI agents independently configure decisions and execute security actions [1]. For example, when a Blue AI agent detects malicious traffic, it can autonomously establish network isolation on affected systems, block outbound connection response, and initiate a forensic investigation without requiring human consent. While accomplishing this level of autonomy requires a significant effort from the defender's part, it also introduces a challenge in terms of implementation, adaptability, and scalability. The combined capabilities can give an advantage to the cyber threat landscape, where intelligent entities must negotiate and compete with their peers in a real time.

In the context of those dynamics between red and blue AI, it is important to examine AI not only from a technical perspective, but also from a system's theoretical and behavioral part. To investigate this evolving scenario much further, we begin to develop conceptual models and simulation frameworks that facilitate experimentation in a controlled environment. These platforms can take many forms, often representing a network, endpoints, traffic flows, and AI agents with varying capabilities. Simulated battlefields [2] allow us to periodically test attack and defense strategies while estimating performance measurements such as detection accuracy, response time, attack success rates, and model drift. The figures provided in this paper illustrate a fundamental abstraction of such environments. It examines the functions and interactions of the Red and Blue AI agents, the battlefield components, and the feedback mechanism through which both agents adapt [2]. Although our architecture is still in a beginning phase version, it provides an opportunity to explore operational dynamics in a broader context of AI cyber conflict.

## III. Related Works

In recent years, there has been an increasing amount of research on the use of Reinforcement Learning (RL) and autonomous agents for Cyber Security. There are a number of essential research topics which demonstrate the potential for RL to be applied to both defensive (in terms of keeping systems safe) and offensive (in developing ways for systems to attack) applications, particularly for hostile situations. This paper shares works focusing on improvements in Autonomous defense, Offensive RL Strategies, Adversarial ML (Machine Learning), Red Teaming Frameworks and Threat Evading Techniques. Instead of focusing primarily on the application of RL technology to develop autonomous systems specifically for Cyber Security defense [3], Palmer et al. (2023) [9] provide a detailed description of how Deep Reinforcement Learning (DRL) is incorporated in the development of Autonomous Cyber defense Agents. Their article is quite extensive, covering many different aspects of Autonomous cyber defense agents (some that relate to how RL technology is applied in the training and development phases), providing an overview of several different methodologies for evaluating how such systems will perform in an ever-changing cyber space environment. In addition, another very important and particularly pertinent

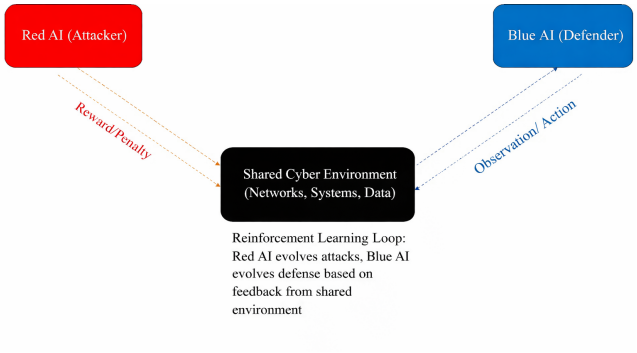


Fig. 1: Adversarial Reinforcement Learning Environment

piece of work discusses features that the authors believe RL agents must have in order to function effectively within a Real-World Autonomous Defense System [3]. The system-level elements that researchers look at include everything from scalability and real-time processing of information to actually securing these systems and allowing them to operate in the real world as AI agents. Academics have produced many new scholarly works that provide clarity on both the potential benefits and risks associated with using RL-based defense systems.

## IV. System Overview

The two autonomous AIs (Red AI/Attacker and Blue AI/Defender) shown in the conceptual framework in Figure 1 describe an active conflict in cyberspace and how both Red AI and Blue AI interact on that shared battlefield through their attack, defense, and knowledge acquisition (exploitation) of attack patterns.

### A. Red AI (attacker)

The capabilities of Red AI agents must not only include offensive capabilities used by opposing forces or completely automated penetration testing solutions. Red AI agents will deploy techniques such as (but are not limited to), automated scanning, the leveraging of known vulnerabilities, machine learning methods to successfully hide their attack efforts from detection and evasion. The primary focus of these agents is to penetrate, disrupt, or otherwise compromise systems' resources while exploiting cyberspace and the associated data transport functions.

### B. Blue AI (Defender)

The blue AI representative is a defensive AI Agent that provides a set of advanced features to defend against adversarial attacks on cyber assets through dynamic detection, Monitoring and Blocking of suspicious activity and through the use of Artificial Intelligence developed based on Adaptive Algorithms [10] for the creation of appropriate Security Policies. The blue AI also provides Predictive Analytical capabilities to assess new or developing threats as they emerge and will take Autonomous Action to mitigate threats to provide availability and maintain the integrity of the System against difficult adversaries [3].

### C. Battlefield in cyberspace:

This is the cyber world where these partnerships or "battlefields" exist, consisting of various networks, applications and other equivalent data. This is what will give Red AI the ability to execute an organized and coordinated attack from the beginning of the battle, whereas Blue AI will use the same environment to create an organized and coordinated counter-attack by creating systems to find, follow, and contain red's attackers and other Red AIs. The feedback loops shown in the diagrams illustrate the continuous flow of data and information between the two AIs on the battlefield and ground level [2]. The ability to learn and modify strategies based on past experiences or current conditions provides both AIs with the ability to make effective and timely decisions. Above all else, the systems that support the AIs will serve as a guide for using AI technology for developing and improving our understanding of how to effectively create more automated and intelligent cyber defense systems [4].

## V. A Conceptual Framework for Simulation

The goal of this research is to create a model of cyber warfare involving entirely autonomous artificial intelligence-based agents; thus, the intention behind developing this model is to enable us to simulate cyber conflict scenarios through the use of autonomous offensive and defensive AIs that can perform independently of each other. The objective is not to model cyber-attacks using predetermined signatures or rules, but rather to observe emergent behaviors where each agent learns, develop, and refine its tactics based on ongoing interactions within the surrounding environment.

By providing autonomous adversarial agents the opportunity to explore, gain access to, defend against and adapt to change within the scope of an environment that replicates the characteristics of a real-world corporate/enterprise information infrastructure, the ultimate goal is to better understand the escalation of cyber conflict between autonomous adversarial agents, develop understanding of the limitations of current AI-based defensive mechanisms, and assess the effectiveness of AI-based defense mechanisms under sustained adaptive cyber-attack. Moreover, this paper aims to investigate the role of feedback loops, reinforcement learning and adversarial modelling in shaping each adversarial agent's tactical responses within the context.

As a conceptual simulation, the efforts of this research are related to form a basis for understanding the dynamics that exist surrounding artificial intelligence cyber interactions and provide insights for the subsequent creation of effective, adaptive, and ethical means for strengthening the security of the computer-based systems. The intention of the conceptual simulation is not to provide explanation by which operational attacks can be developed and implemented, nor is the objective to identify and describe how the framework will deal with scenarios such as exploitations or vulnerabilities; the intent is to form a research infrastructure that will facilitate and increase the knowledge regarding how autonomous systems react and adapt.

## VI. Cyberspace Battlefield: The Domain of Conflict

A virtual battlefield exists within the broader picture of cyberspace. The red AI represents all offensive actors while the blue AI is on the defensive side. Furthermore, cyberspace itself goes beyond just being an abstract concept. Instead, it consists of the environments that serve as the foundation for all real cyber action, such as: clouds, corporate networks, applications [11], and data flows. As such, knowing how the virtual battlefield works will help in developing and designing cyber strategies that use AI as an integral part of their operation.

It is required that agents in cyberspace possess the following attributes: an awareness of context (i.e., knowledge of what is happening around them), adaptability (i.e., ability to change in response to changes in their environment), and the readiness to learn (i.e., learn from mistakes and successes). However, because of the challenges involved in cyberspace, various defense mechanisms must work together synergistically to defeat attacks that increasingly rely on automated and intelligent means.

To achieve the goal of creating tools, simulating the most realistic engagements, and designing cyber defense systems [4] that will be resilient, self-governing, and operate successfully in a challenging and contested digital environment, we must collectively define or articulate these aspects of structure, interaction with one another, and the development of strategies regarding one or more of the above.

The complexity of simulating Cyberspace conflict environments is so great that the processes for creating simulations comparable to other types of "highly disruptive" environments are not adequate.

Just as modelling and understanding Cyberspace as a battlefield is vital to developing a scientifically-valid foundation for AI-based Cybersecurity solutions, such simulation will also assist us in creating greater resilience for our Digital Infrastructure against newer and more diverse types of threats.

To summarize, Cyberspace as a battlefield continues to be both highly dynamic and fluid, in addition to being borderless and static. There are many examples of Autonomous Agents not only keeping up with the pace of operations, but also completing their own Adaptive Operations, with either remotely-sourced or live-generated information.

## VII. Conflict Dynamics in AI-Driven Cyber Engagements

In Figure 2 we can see two different AI agents attacking and defending themselves against each other in an environment set up to mimic a battlefield situation. After every time the agents make their moves, they will both have the chance to adjust their methods and rules and will know based upon past experiences. The first attack made by an attacker will begin the process of actively searching for vulnerabilities in a target. This step will also be based upon the success or failure of previous attacks.

The first attack made will be documented under an adversary characteristic module. The module will document all of the tools that have been used by the attacker and the order in which they were used, as well as what their final goal was and any unusual behaviors that were detected. This module can be accessed and modified by both sides and provides information about how the attackers' plans changed over time. Once the module has been updated, the threat intelligence system [10] accesses this new information and adds it to the existing threat intelligence logs and past events. A threat intelligence system provides monitoring and guidance to the defender by identifying how their protection methods need to be adjusted, providing suggestions about what the attacker will likely do next and providing information about any new patterns of threats.

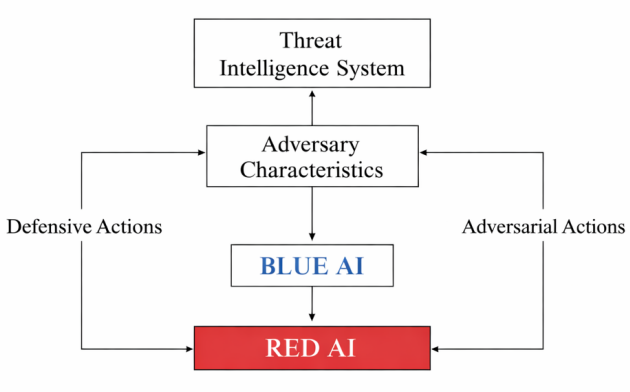


Fig. 2: System's Proposed Mechanism

A defensive artificial intelligence (AI) system utilizes collected information to implement countermeasures, deploy deceptions, and utilize resilience rules. The system self-learns through reinforcement and adapts its responses based upon the opponent's change of tactics (from success to failure) as well as its prior assaults.

Figure 2 shows the autonomous adversarial mechanism as indicative of an ongoing battle without any predetermined scripts by either agent(s). Both agents are autonomous, and also both agents learn from each other as they engage in the battle. Despite the fact that agents can train to execute cyber-attacks, the diagram accurately illustrates the relationship between cognitive cyber dueling. In cognitive cyber duels, the agents' primary goal is more than simply to defend and attack, but rather, to observe how intelligent agents survive and adapt to conflicting countermeasures by developing refined strategies with an ultimate goal being the domination of a given arena. The system provides a comprehensive means to evaluate resilience, provide feedback with regard to adaptive defensive measures, and assess the ethical considerations associated with AI-driven conflict.

The Defensive Artificial Intelligence System uses the data it collects to provide countermeasures, misinformation to attackers and follow predefined resilience guidelines.

Through reinforcement learning [1], it develops its own strategy based on how successful its earlier behaviors were versus how it plans to respond to the attacker's latest attempt. The moves taken by the defender also provide examples of how to train the attacker; therefore, both the defender and attacker incorporate the behaviors of the opposite side in their

training data to improve their operational capabilities.

As a result, both players are able to provide useful training data for the other.

While this diagram serves as a basis for researching how both intelligent agents behave under pressure, develop new adaptive responses through observation of their opponent, as well as hone their competitive strategy for attacking and defending, it can also be used to establish resilience assessment, adaptive defense assessment and ethical parameters for AI conflict. This interdependency ensures continuous adaptation, as detailed in the feedback mechanisms of Section VII.

TABLE I: System's Core Dependency Mapping

| **Source Component** | **Target Component** | **Feedback Type** |
|---|---|---|
| Red AI System | Adversary Characteristics | Behavioral |
| Blue AI System | Threat Intelligence System | Action |
| Threat Intelligence System | Defensive AI System | Direction |
| Adversary Characteristics | Threat Intelligence System | Profile Input |
| Defensive Actions | Adversary Characteristics | Countermeasure |

Table 1 illustrates the interplay between the various elements that comprise AI-Cyber Systems through the relationships that exist in feedback loops amongst multiple devices and systems.

At the core of the feedback loops of all components within this ecosystem of intelligent systems lies the Adversarial Artificial Intelligence System (AI System), which generates offensive strategies based on adversarial tactics, and generates dynamic characteristics of each adversary (Dynamic Profiling), which measure the adaptability and fluidity of enemy tactics through time.

The dynamic profile generated by the Adversarial AI System is then used by the Threat Intelligence System to integrate all gathered data and analyze the dynamic profiles of the adversaries to create predictive insight reports.

This predictive insight report serves as the blueprint for the Defensive Artificial Intelligence System's (AI System) immediate countermeasure and resilience methodologies [10].

Each of the Defensive Actions that have been carried out as a result of using this system will not only eliminate Cyber Threat Vendors but also change the characteristics that drove the Adversarial AI System's Dynamic Profile and create new Tactics that the Adversarial AI System's path has produced.

Through the combination of the output from one system being the input of a second system and vice versa, the components develop into an ever-evolving exchange of adaptive traits within a mutually-supportive ecosystem.

Interdependencies among all modules of the model ensures that no single module operates independently. Each module has a strategic influence on every other module, enhancing the cognitive complexity and realism of the simulation.

This model also serves as a core concept for modelling autonomous cyber conflicts, wherein agents can continually learn, adapt and alter the virtual environment based upon the actions of other agents engaged in the conflict.

## VIII. Co-Evolutionary Dynamics of Red and Blue AI:

Basically, co-evolving systems are a back-and-forth battle between a Red AI which is constantly creating new methods of attacking the system (for example, by the use of zero-day exploits and adversarial integrations) and a Blue AI that is becoming more powerful through the means of using various of the latest AI approaches (e.g., Anomaly detection, Reinforcement learning [1] and Adversarial training) to combat the Red AI. For both the Red and the Blue AI, the ability to have machine learning based models will allow them to take part in this co-evolution [7]. Like the Blue AI, Red AI also has the ability to learn algorithms and use the reinforcement learning capability [9] in order to find the best method for exploiting a vulnerability, thus enabling Red AI to evade detection as well as offering a way to blend in with "normal" behavior. In addition, the Blue AI must remain alerted in watching for new attack points and anticipating any threats that could develop both agents during and after the attack. One of the main features of this type of co-evolution is that the AI agents on each side will have the tendency to defend against not only the majority of threats developed, but to proactively guard against the spread of these threats based on what the AI has learned during its training. The evolution of AI agents will also adequately prepare them for any subsequent or future attacks regardless of how their original models performed. It is very important that both Blue and Red AI use continuous learning/self-adapt strategy so that both of these Agents can respond and adjust to their adversary's changing behavior. By utilizing these strategies, both Agents will be able to have a real-time, real-world interaction in which they are continuously learning, adjusting, and responding based on the information they gather from each other. This gives us an insight into how effective automated Red Teaming [5] Blue Teaming enables both Agents to perform effectively when they work together.

As RED AI evolves, it impacts the workload placed upon BLUE AI. Similarly, as BLUE AI evolves, it impacts the workload placed upon RED AI. Therefore, by implementing agent-based modeling and reinforcement learning [1] of both RED and BLUE AIs, the AIs will be able to create a simulation of this EVOLUTIONARY CO-EVOLUTIONARY PROCESS whereby Agents will continue to adjust their learning curve from the "Trial and Error" process until they attain their maximum performance capabilities. The Adversary Characteristics of both RED and BLUE AI allow AIs to not only receive and analyze information about what actions the other is taking, but also how the outcome of the coevolutionary development of the AI agents evolve response behaviors based upon understanding the similarities between both AIs.

As with most systems, there is an ever-growing level of computer autonomy that can cause both developed and evolving AI agents to take coevolutionary aggressive actions without the knowledge of the human involvement. For example, by identifying the different impact vectors for the AI agents' actions, the red AI may identify how to develop an event scenario that resulted in undesired and unexpected operating errors. The blue AI agent could (in its own evolution) to impose the limitation and the degradation of the available operational systems, thus, resulting in reduced operational processes including denial of service events.

It is of the utmost importance to develop an understanding

of the coevolutionary dynamics in terms of AI development from the red AIs perspective to the blue AIs perspective, and to understand how evolved AI agents can respond to pressure, escalation, and robust adaptive and operationalized AI agents' configurations to accomplish human objectives by developing methodologies that measure how rapidly the evolving threat may have impacted the heterogeneous systems after an incident.

With the Red and Blue AIs both growing together via many feedback loops in interacting with each other and via making changes in the physical environment (cyberspace) in which they operate, the manner in which both AIs evolve depends upon the output and feedback from the outcome of all attacks and defenses, thereby providing each AI with direct feedback for making changes to its model through iterative learning. Specifically, Red AI will evolve more sophisticated methods for avoiding detection, using vulnerabilities, and falsifying data input. Conversely, Blue AI will increase its ability to detect attacks, improve its response strategies, and enhance its resilience through various strategies such as adversarial training and anomaly detection [18].

The Red AI has evolved from being a tool designed to check for known vulnerabilities to being characterized as an "autonomous explorer" of an attack surface through the use of reinforcement learning techniques in order to identify and exploit and create malicious polymorphic malware that changes its behavior with each new attack to other agent's system. On the other hand, the ability of the Blue AI to utilize sophisticated anomaly detection algorithms to identify deviations in the conflict process will assist in reducing the time when an attack happens and when a defensive action occurs in response [19]. The combined use of Red AI and Blue AI will create a self-reinforcing closed loop relationship between the two agents where any action taken by one agent generates an immediate response from the other [15].

The continual expansion of this conflict is related to the continuous competitive learning and adaptation cycle as the two systems are trying to accomplish increasingly complex tasks. For instance, Red AI has a much larger number of ways to exploit Blue AI's vulnerabilities as compared with Blue AI's ability to exploit Red AI's attacking methods. By attacking Blue AI where there are deficiencies (vulnerabilities) in detection and response capabilities, Red AI can get enough information from Blue AI especially on how Red AI can affect Blue AI. We are talking here on the method of the attack, the logs that were obtained during the attack, and any other credentials or other related data that were obtained during the conflict process [16]. This will thereby result in improved training data for the Blue AI side, which in turn will get to know how Red AI's behave. This will on the other hand will enhances Red AI's ability to detect future similar types of anomalies (especially during real-time incident detection). As Red AI continues to improve Blue AI's real-time incident detection capabilities through the previous discussion, Red AI will also continue to create new attack techniques against Blue AI. The techniques learned by Blue AI can directly make it much harder for Red AI as each time Blue AI improves its capabilities (i.e., as in how well it accurately detects false positives), It will have the capabilities in requiring a higher threshold for detecting the same type of anomalies in a real time manner. Hence, the complexity that is made to the Red AI will lead to an increase consumption in the resources required to conduct an attack, thus forcing it to adopt different attack techniques or even utilizing stealthy methods to avoid being detected.

Both AI agents Red and Blue, can learn from each other by receiving rewards for their behavior during the reinforcement learning process [17]. The goal of Red is to optimize its behavior so that it compromises the target while avoiding being detected. On the other hand, Blue optimizes its behavior to detect a compromise behavior as accurate as possible while minimizing false positives detection results. By using this process of trial and error, the two agents can be able to find their way gradually towards the best secure environment that cannot be possibly reached by traditional human methods. Yet, the autonomous nature of these systems poses a very serious risk in terms of network security, as the agents attempt to optimize their performance without ethical constraints. If this were to occur, there is a possibility of a "runaway escalation" occurring between those two agents. For instance, Red can take advantage of its reward function to find a way to cause a complete failure of the system simply by creating a DoS attack exploiting a vulnerability on Blue AI's vulnerability as this is considered an effective way to satisfy the conditions of its reward implemented functions. At the end, there will be a limit on the agent's ability to be autonomous. There will be a need for human oversight when there are such incidents and there must be strong validation procedures to ensure that the evolving agent strategies are aligned to enhance organizational business objectives as opposed to simply gaining artificial rewards.

## IX. Autonomous Red-Blue Agent Loop for Adaptive Cyber Defense

The architecture is being modeled as a cognitive duel among autonomous agents using a closed-loop framework that has been developed for the research design and simulation of our system. The framework has four components that are interconnected: the Threat Intelligence System, the profiles of the Adversaries that are being simulated, and the Red and Blue Agents (AI) that emulate the Adversary — in our case the attacker (RED) and the Defender (BLUE).

The Red Agent utilizes reinforcement learning methods and a reward mechanism based on the successful exploitation of targets, persistence, and ability to evade detection, whereas the Blue Agent has been developed using both reinforcement Learning and supervised learning in order to improve its ability to Detect attacks targeting it, contain threats as quickly as possible, and adapt to emerging threats.

Both agents operate within a controlled simulated environment to create a realistic representation of normal network conditions (e.g., endpoints, vulnerabilities, and monitoring systems). The Threat Intelligence System then reviews each duel and updates the Profile of Threat(s), allowing it to refine the Blue Agent's Defenses and, thus, form a self-sustaining loop of ongoing evolution.

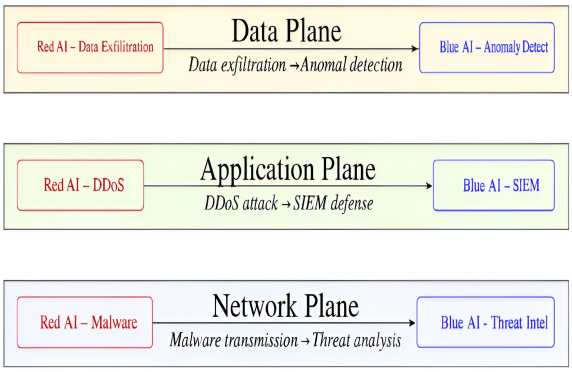


Fig. 3: Proposed System's Layers

Both of the agents provide a virtual representation of an actual network in terms of the endpoints, vulnerabilities, telemetry/monitoring infrastructure, etc. and can simulate all aspects of an interaction in a safe manner without putting an organization at risk or having the costs associated with a real-world network by utilizing simulated environments. These same simulated environments also allowed us to track emergent behaviors, escalation/response patterns, and cycles of adaptation through various phases of the adversarial cyber kill chain (Reconnaissance, Weatherization, Delivery, Exploitation, Installation, Command and Control (C2) and actions on objectives) throughout the duration of the agent's adversarial interaction.

Every cycle of iteration; the Threat Intelligence System receives the responses from both of the agents (attack signature(s), behavior pattern(s), failed exploit chain(s), misclassification of defensive networks, etc.) and creates a new Threat Profile based upon those data. The Threat Profile is a representation of how the components of the adversary are evolving, and it also serves as a model for future defensive strategies for the Blue Agent. The process of both offensive and defensive AI models working together creates a co-evolving, self-perpetuating cycle; in which each are continuing to adapt over time.

There should he Multiple rounds of the performance evaluations of the system to be conducted to determine any pitfalls of the system through an evaluation of the following measurements: Attack Success Rate, Speed of Adaptation to Defensive Response, Cognitive Drift Index, Accuracy of Threat Intelligence, and Feedback Latency.

## X. AI-Driven Cyber Conflict Across Infrastructure Layers

The system described in Figure 3 shows how it will respond to Cyberattacks through a multi-layer interaction with both Red AI (Offensive) and Blue AI (Defensive), across three layers of plane-type Architecture: Data Plane, Application Plane, and Network Plane. Red AI collects and retrieves information from the Data Plane Layer through Data Exfiltration targeting Databases, Database Storage, and Data Traffic. Blue AI will utilize an Anomaly Detection Technique in the Data Plane Layer to detect anomaly usage patterns that could indicate any type of anomalous/unsanctioned threat or a breach to Sensitive Data. Moving up to the Application Plane, Red AI is launching Distributed Denial of Service (DDoS) Attack against Application Containerized Web Services (APIs) for the purposes of compromising and disrupting operations. In response to Red AI's attack, Blue AI is responding with Security Information/Event Management (SIEM) solutions in the Application Plane to provide log and event data visibility and intervention capabilities to enable prompt and thorough analysis of logs and timely identification of developing threats. In the Network Plane, Red AI propagates Malware through targeting compromised routers, FIREWALL appliances, and data traffic flows while Blue AI utilizes Threat Intelligence to anticipate, detect, and mitigate Intrusions. These different levels represent a constantly changing environment; within this environment, the two sides (the defenders and offenders) constantly alter their strategies and tactics of cognitive cyber-warfare on progressively increasing levels of sophistication. As a result, a sophisticated and adaptive counter-measure system will be the main outcome to protect a organization's cyber ecosystem.

## XI. Assumptions

Several assumptions are presented here to narrow the scope of the research in order to retain control of Red AI and Blue AI complexities. Initially, it is assumed that both agents act independently, performing tasks and making decisions without direct involvement of human agents. The Red AI agent is considered to operate as a highly advanced attacker who is trying to exploit vulnerabilities in a particular system [2]. The offense actions of the Red AI agent can be seen as following typical cyberattack tactics in the cyberspace that involve the following exploitations tactics. For instance, the Red AI agent may conduct reconnaissance, identify some vulnerabilities, exploit those vulnerabilities, or even use adversarial machine learning, but in the case of lacking enough knowledge of the system's internal and essential specifications; the performance of system could be inaccurate.

On the other end, the Blue AI agent is considered to be a system that constantly reacts to threat detections, acts upon it, and monitors system behavior [2]. This entails making use of detection mechanisms, filtering rules, or learned defense strategies. In this regard, there should be one of the factors obtained by a Blue AI agent to observe anomalies and taking the equivalent decisions while not only depending solely on predefined signatures/rules.

Because AI Technologies continues to become more sophisticated, it is anticipated that adversaries will utilize Artificial Intelligence for decision-making, pattern recognition and adaptive techniques, which will be comparable to or may exceed human cognitive abilities. Autonomous adversaries will evolve rapidly due to defenses [4] put in place by cyber security systems.

Also, this research is based on the premise that the AI-driven adversaries will adapt and improve their tactics by learning from their experience of engaging with a security system, just as a human attacker would. Consequently, it is

reasonable to assume that AI-driven adversaries will be able to learn and experience in their interactions to change their attack modes to prevent themselves from being detected or gaining an unauthorized access [8].

The cyberspace battlefield is simply an abstract simulation environment. It represents various digital assets: networks, applications, data flow, and so on. Access to networks and data in this cyberspace battlefield does not obey real-world network configurations; it rather serves as a flexible arena to evaluate interactions and responses of the presented agents. The resulting actions will create sufficiently outstanding feedback signals within the system to facilitate the learning and adaptation of both agents.

## XII. Challenges and Considerations

There are many considerations and challenges to take into account in regards to designing and simulating autonomous interactions Red AI and Blue AI agents in the cyberspace environment. Even though we are assuming that the system will provide various types of flexibility and clarity, we must also consider the limits and restrictions posed by applying the mechanism as in the proposed system's layers discussed in the figure above. One of the most significant challenges with successfully simulating the detections of autonomous attacker (Red AI) and defender (Blue AI) agents is creating realistic behaviors for both of them. In the real world, attackers use offensive tactics that can be incredibly complex and dynamic based on the nature of the organization's environment. Managing this degree of heterogeneity in regards to simulating Red AI responses will require not only carefully constructing sets of actions that defined the attacks but an equally developed ability incorporated into the Red AI's behaviors as such actions would involve learning patterns of activities and adapting themselves accordingly. Similarly, in order to create useful, realistic responses for our Blue AI agent; the need to implement an advanced defender actions capable of deception, predictive detection, dynamic resource determination and policy alteration rather than developing predefined static responses are required. Another challenge that is the fact that autonomous adversaries are growing rapidly in a way considered to be difficult to anticipate. This requires the Blue AI systems to find ways to adapt to the evolving attacks challenges without sacrificing the system's performance.

This will lead us to another challenge; which is creating a realistic simulation environment that accurately reflect the operations happening within the cyberspace. There is a high chance caused by implementing a poor simulation that can lead to misleading results. This will affect the resilience [10] and efficiency of the implemented system especially when adapting with current or new threats.

Moreover; Integrating the proposed AI agents with the current cybersecurity ecosystem (e.g., IDS/IPS and SIEMsrequire a robust APIs and data normalization. Without proper integration of the current cybersecurity tools and components; the real time response and threat intelligence will be affected and may be compromised especially during the data sharing process.

## XIII. Discussion

The way autonomous offensive AI interacts with its counterpart autonomous defensive AI [3] demonstrates the existence of a rapidly accelerating race between who can create better cyber-security systems. The attack and defense terms are no longer considered separate issues, instead, they are operating as a one continuous loop where both offensive and defensive agents are reacting and observing each other in a real-time manner. Additionally, by combining Threat Intelligence and Adversarial Characteristics into the architecture, we highlighted the effectiveness of using a dynamic threat towards the preparation of future threats, as it allows for constant evolution for each component of Cyberspace.

The three-layer architecture has demonstrated how an adversarial loop evolves through adversarial strategy changes, in which offensive and defensive agents interact with one another. The action and reaction of each party provides an evolving set of resilience tactics for defensive agents. Ultimately, this combination of agents will provide an environment that reflects the closed-loop learning environments within which the next generation of threats is being developed and tested.

In this research we examined both the capability of applying resilience [10] and/or adapting Cyber Security Systems and then implementing this new approach to strengthen the cybersecurity ecosystem by showing how our defense-in-depth layering strategy can aid in mitigating the threat that new, rapidly advancing technologies use to target our existing defense systems.

Over time, the current traditional defense systems will be hindered by their own nature to continue providing a means of protection against new cyber threats. Due to the nature of traditional defending systems being reactive and not resilient to new attack methods or new emerging technologies; this will continue to deliver a constant cycle of cybersecurity products that provide no real solution in protecting against futuristic and more complicated cyber threats.

Given the complex nature of the cyber threat landscape, the three-layered system that we proposed earlier represents a very new and innovative way to combat network attacks. Implementation of our proposed system not only helps network analysts identifying new exploit techniques utilized by adversaries but provides valuable insight into how adversaries will generally behave within the cyberspace arena. Thus, providing a structured environment for experimentation on how cyber security systems could effectively counter Intelligent Autonomous Adversaries.

Although we understand that there are many challenges to implementing this new paradigm, we believe that this will be a very promising and interesting topic of further research. One aspect of the research would be expanding our design into a Multi-Agent Environment; if we added several existing Cybersecurity tools and components to our design, this would allow the entire Cybersecurity Environment to be better prepared to defend against more sophisticated autonomous cyber threats.

Additionally, we view the AI-driven Battles approach as not just a new technical exercise; but rather as a new strategic paradigm through which we can determine how we will

Protect our organizations and individuals from future cybersecurity threats.

## XIV. Conclusion

Autonomous cyber warfare can be demonstrated in an generalized way through examining how Red AI (attacker) and Blue AI (Defender) interact in a Battlefield-Cyber environment. This paper creates a simulated environment in order to understand how an automated cyber warfare will behave, and how Adaptive Defenses will function as agents in a cyberwar Battlefield. The key to eliminating the complexity of global cyber infrastructures is through creating a conceptual framework on how the Proposed Cyber AI System will work, in particular how Red AI will use a reward loop for launching attacks and how Blue AI generates defense responses in a real-time manner. The main contribution of this Study is to propose a new Architecture Model for Adversarial AI Simulation that includes Assumptions, Challenges and Limitations of Agent Interactions. Based on how both the Red and Blue AI can evolve, adapt and respond to attacks in real-time, Future cyber security Solutions could be designed to be smarter, scalable, and automatically deployed, with the added potential for multiple levels of Protection. In this research, we presented a simple conceptual architecture as our promising mechanism of how is our vision to the AI adversarial agents will be beneficial, but we will continue to investigate more as future work on robust models such as graph neural networks that will add efficiency to the presented architecture and thus strengthen the network environment security. Also, this research presents a conceptual mechanism for developing a simulator; therefore, it does not provide a detailed methodology on such implementation, nor does it present any test results in regards to this matter, as this task will continue to be executed as a future work in subsequent phases.

## Acknowledgment

I would like to thank Professor Hiroki Takakura for his expert advice and encouragement throughout this research paper.

## References

[1] S. Castro, R. Campbell, N. Villalobos, J Duan and A. Cardenas, "Large Language Models are Autonomous Cyber Defenders," *arXiv preprint*, arXiv:2505.04843, 2025. [Online]. Available: https://arxiv.org/abs/2505.04843

[2] DARPA, "Securing Artificial Intelligence for Battlefield Effective Robustness (SABER)," Defense Advanced Research Projects Agency, 2025. [Online]. Available: https://www.darpa.mil/sites/default/files/attachment/2025-03/program-darpa-saber-proposer-day-presentation.pdf

[3] M. Foley, C. Hicks, K. Highnamand and V. Mavroudis, "Autonomous Network Defence Using Reinforcement Learning," *arXiv preprint*, arXiv:2409.18197, 2024. [Online]. Available: https://arxiv.org/abs/2409.18197

[4] Y. Han, D. Hubczenko, P. Montague, O. De Vel, T. Rubinstein, C. Leckie, T. Alpcan, and S. Erfani, "Adversarial Reinforcement Learning under Partial Observability in Autonomous Computer Network Defence," *arXiv preprint*, arXiv:1902.09062, 2019. [Online]. Available: https://arxiv.org/abs/1902.09062

[5] S. Majumdar, B. Pendleton, and A. Gupta, "Red Teaming AI Red Teaming," *arXiv preprint*, arXiv:2507.05538, 2025. [Online]. Available: https://arxiv.org/pdf/2507.05538

[6] A. Lohn, A. Knack, and A. Jackson, "Autonomous Cyber Defence: A roadmap from lab to ops," Centre for Emerging Technology and Security (CSET), 2023. [Online]. Available: https://cetas.turing.ac.uk/sites/default/files/2023-06/autonomous_cyber_defence_final_report.pdf

[7] A. Vassilev, A. Oprea, Alie. Fordyce, H. Anderson, X. Davies and M. Hamin, National Institute of Standards and Technology (NIST), "Adversarial Machine Learning A Taxonomy and Terminology of Attacks and Mitigations (NIST.AI.100-2e2025)," 2025. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2025.pdf

[8] Outflank (K. Avery), "AI-Powered Malware Evades Microsoft Defender Security Checks Approximately 8% of the Time," *Windows Central*, Jul. 2025. [Online]. Available: https://www.windowscentral.com/artificial-intelligence/ai-powered-malware-eludes-microsoft-defenders-security-checks-8-percent

[9] G. Palmer, C. Parry, D. Harrold and C.Willis, "Deep Reinforcement Learning for Autonomous Cyber Defence: A Survey," *arXiv preprint*, arXiv:2310.07745, 2023. [Online]. Available: https://arxiv.org/abs/2310.07745

[10] F. Hernandez, "AI vs. AI: The Evolution of Offensive and Defensive AI Techniques in Cybersecurity," *TechRxiv preprint*, 2025. [Online]. Available: https://doi.org/10.36227/techrxiv.173937772.29983104/v1

[11] K. N. Kseniia and A. Minbaleev, "Legal Support of Cybersecurity in the Field of Application of Artificial Intelligence Technology," *2020 International Conference on Quality Management, Transport and Information Security, Information Technologies (IT&QM&IS)*, pp. 59–62. doi: 10.1109/ITQMIS51053.2020.9322905.

[12] K. Y. Nikolskaia and V. B. Naumov, "The Relationship between Cybersecurity and Artificial Intelligence," *2021 International Conference on Quality Management, Transport and Information Security, Information Technologies (IT&QM&IS)*, 2021, pp. 94–97. doi: 10.1109/ITQMIS53292.2021.9642782.

[13] G. Liu, H. Wan and L. Zhang, "Application of Artificial Intelligence in Computer Network Technology in big data era," *2021 2nd International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT)*, 2021, pp. 687–690. doi: 10.1109/AINIT54228.2021.00139.

[14] D. Rosch-Grace and J. Straub, "Considering the Implications of Artificial Intelligence, Quantum Computing, and Cybersecurity ," *2022 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2022, pp. 1080–1082. doi: 10.1109/CSCI58124.2022.00191.

[15] K. Mahmood, E. Rathbun, R. Sahu, M. Van Dijk, S. Ahmad, and C. Ding, "Game Theoretic Mixed Experts for Combinational Adversarial Machine Learning," *IEEE Access*, vol. 13, pp. 158887–158905, 2025, doi: 10.1109/ACCESS.2025.3608117.

[16] T. A. Khaleel, "Developing robust machine learning models to defend against adversarial attacks in the field of cybersecurity," *2024 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, Istanbul, Turkiye, 2024, pp. 1–7, doi: 10.1109/HORA61326.2024.10550799.

[17] Y. Wang, M. Liu, J. Chen, and H. Zhang, "Adversarial Attacks and Defenses in Machine Learning-Empowered Communication Systems and Networks: A Contemporary Survey," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 4, pp. 2245–2298, Fourthquarter 2023, doi: 10.1109/COMST.2023.3319492.

[18] B. K. Sharma, A. K. Rai, P. Kumar, A. K. Rai, and K. Tripathi, "Hybrid Models for Effective Adversarial Attack Detection in Cyberspace Using Machine Learning," *2025 3rd International Conference on Disruptive Technologies (ICDT)*, Greater Noida, India, 2025, pp. 1308–1313, doi: 10.1109/ICDT63985.2025.10986745.

[19] Z. He, D. Davila, S. Bi, T. Wang, and T. Hou, "Machine Learning for Cybersecurity: A Survey of Applications, Adversarial Challenges, and Future Research Directions," *Electronics*, vol. 14, no. 23, pp. 4563, 2025, doi: 10.3390/electronics14234563.

**Nader Shahata** Having a Bachelor Degree from King Abdul-Aziz University on Computer Science in 2008. Received a Master Degree from Canberra University on Information Technology and Systems in 2012. Currently; I am a project researcher in the National Institute of Informatics.