Artificial Intelligence for Cyber-Security: A Deceive to Defend Concept

Tarushi Jain
Department of Information Technology, NIT Srinagar,
Jammu and Kashmir, India
tarushijain2008@gmail.com

Saniya Zahoor
Department of Information Technology, NIT Srinagar,
Jammu and Kashmir, India
saniyazahoor@nitsri.ac.in

Shabir Ahmad Sofi
Department of Information Technology, NIT Srinagar
Jammu and Kashmir, India
shabir@nitsri.ac.in

***Abstract*— This paper introduces a framework called Artificial Intelligence for Cyber-Security: A Deceive to Defend Concept which revisits the integration of AI as an aspect of securing digital environments. The recognize Deceive to Defend model includes predictive analytics, real time threat detection and response, autonomous response, in addition to active deception strategies to proactively manipulate, mislead, and monitor adversaries. We maintain that cyber deception must be recognized as an important consideration for defender-centric security mechanisms rather than incidental. We integrate recent research that shows how an AI-enabled deception strategy allows the defender to isolate the attackers, increases defender situational awareness and ultimately allows for the defender to have the power differential. Our paper advances the conversation to think about the ethical, the legal, and morality issues surrounding operationalizing deception; and consider deception as a method not only to protect each defender as an incident occurs, but engaging in deterrence. We illustrate here the concept in the deception to defend concept as a working concept for a deception model that includes deception as an important AI-enabled mechanism for defense, submitting it is the next frontier in technology in cyber security mechanisms.**

Keywords—Artificial Intelligence, Cyber-security, deception

I. INTRODUCTION

Today's world, with its global nature, has created a diverse and ever-evolving threat environment in cyberspace. Cyber-attacks are no longer limited to trading actions or viruses, and now can include advanced persistent threats (APTs), zero-day vulnerabilities, ransomware-as-a-service and adversarial AI. Current cyber-security methods that are rule and/or response based cannot compete with the complexity, diversity, scale and speed of attacks today [1]. AI presents a novel variable that can lessen attacks due to the ability to rapidly learn about normal behavior from large data sets, predict when a threat will occur, and respond autonomously to a threat. AI has already changed detection and response capabilities and will continue to operationalize those capabilities to support scalable, flexible, proactive defense [2].

The current changes in cyber security demonstrate a move away from classic, deterministic, rules-based defenses to intelligent, adaptive, and AI-enabled solutions. We have never seen this level of sophistication of cyber threats ever before. In addition to scope and scale of threats and vulnerabilities, we have APTs, malicious actor adversarial AI, zero-day vulnerabilities, and ransom-as-a-service providers. Therefore, traditional solutions such as firewalls to anti-virus systems to static solutions are simply inadequate for a dynamic threat environment. In addition, AI has resulted in what is being referred to as a revolution, with capabilities being provided for integrating analysis of large-data, anomaly detection in real-time, and automating responses, to mention a few. Cyber deception technology as an enhancement to an AI defined strategy can offer distinct value, and extend the capabilities in question; detection of intruders, deceiving or misdirecting them, observing their behaviors, and defending relying on realistic decoys and flexible environments. The "Deceive to Defend" model, it can be argued, is creating a new definitional category of cyber security by using predictive threat capabilities to not just deceive adversaries, but, gather intelligence and entrap adversaries as well, all apart of a simple model to outsmart. Figure 1 shows sequentially what cyber-security has evolved to in greater complexity and response capabilities.

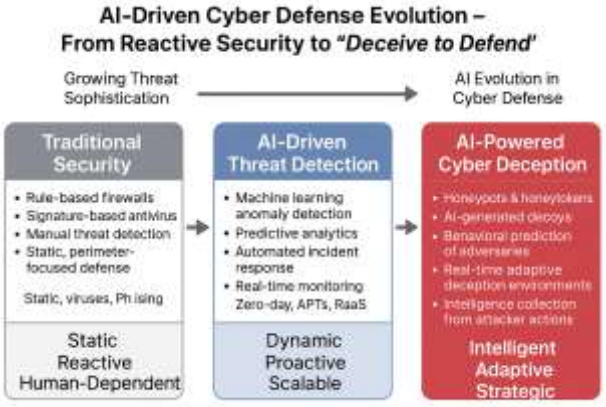

FIGURE 1: AI-DRIVEN CYBER DEFENCE EVOLUTION

Conventional cybersecurity has depended on static approaches to deploying countermeasure tactics, such as, rule-based Firewalls and Malware solutions that defend against common threats but are deficient against modern attack vectors. Dynamic and real-time capabilities such as AI-enabled threat detection create enhanced anomaly detection, automate actions, and build corresponding scaled defenses. The expansive capabilities of this new model can also be found in advanced capabilities of distraction and deception with the translation into intelligent systems of realistic decoys (ex. honeypots, honey tokens) to predict actions and adapt in real time. This next-level evolutionary strategic development is "Distraction to Defend," as, the development of adaptive countermeasures, to mislead, entrap, and better understand today's evolving cyber adversaries to enhance posture.

Nonetheless, the artificial intelligence-enabled Defensive Cyber operations involves a new area, not just with detection, or response, but cyber deception which implies the use of confusion, misdirection, or trapping cyber adversaries, same way the use of AI in deception technology can provide for real-world decoy environments, it can use anticipated attacker behavior before any exploitation occurs, and it can manipulate the defense environment while it occurs. Deception is much different from the previous notion of a static security. It is in this instance that we introduce the "Deceive to Defend" concept: a conceptual framework for a practical and strategic measure to increased cybersecurity posture with the use of cybersecurity deception aligned with AI. The "deceive to defend" model provides the opportunity to utilize AI-driven deception as a methodology for misleading, trapping, and collecting intelligence on malicious actors. Thus, with cyber deception embedded into AI systems, the cyber environment can be transformed to create an environment that provides the opportunity to study the attack, stop the attack, and study the attacker. The paper is then organized in the following way, addressing the theoretical foundations in deception-based security, real-world applications, limitations, and ethical concerns surrounding cyber deception [3].

The paper is organized as follows: Section II discusses the literature survey, Section III presents the proposed CDDS based on Deception as a Defense Strategy, Section IV covers Challenges and Ethical Considerations, and Section V provides conclusions.

II. Literature Survey

Artificial Intelligence has quickly become a key part of modern cyber-security practices. Many studies look into how to use it for threat detection, defense without human intervention, deception, and strategic counterintelligence. This section shares insights from ten important papers in the field.

In [4], the authors provide a thorough review of how AI can be leveraged for cyber-security countermeasures. They discuss the application of machine learning (ML), deep learning (DL), and neural networks in intrusion detection, malware detection, and anomaly detection. They also discuss the potential of employing AI to monitor large-scale traffic and deal with zero-day attacks. In [5], the authors study the vulnerabilities of Industrial IoT systems and propose a method for finding new malware patterns, based on statistical drift. The detection accuracy for their case studies reached 95.2% with an F1-score of 94%, highlighting that adaptive ML classifiers outperformed static models in dynamic threat environments. In [6], the authors present an AI to model the cyber-security for a smart city based on the Complex Adaptive Systems (CAS) framework, as well as theories of behavior regarding attitudes towards technology (i.e. Technology Acceptance Model (TAM), Technology of Planned Behavior (TPB)). This model can be used to improve awareness and resilience to threats in multi-layered IoT systems.

Authors in [7] introduce the MDATA-based ACAM model to identify multi-step cyberattacks. The authors argue against the traditional rule-based IDS systems, and how by combining temporal (time-related) and spatial (location-related) data through AI modules could dismiss false positives and gain an overall improved performance in identifying coordinated attacks. Authors in [8] investigate AI-enabled malware and intrusion detection for contemporary digital infrastructure. It utilizes several ML classifiers and ensemble models for anomaly detection, with little exploration into deployment and systems connection in vast networks. The authors of [9] studied deception enabled by AI. They describe how adversarial machine learning could be employed to dynamically design deceptive decoys that could be believable. They maintain that deception enabled by AI can mislead attackers as well as provide intelligence as real-time behavior patterns are analyzed. Authors in [10] are open to this ethical turn to study governance/ regulatory issues surrounding deception enabled by AI. Bonfanti et al. further consider the potential for ethical use of an AI-enabled strategy with deception that would require principles, enforcement, and policies.

Authors in [11] examine honeypots that are supported using AI and explore the way machine learning can change deceptive responses on behalf of the deceived based on the behavior utilized by attackers. The authors conclude that deception-based technologies provide an improved ability to entrap adversaries while adding resiliency to cyber defense strategies. Authors in [12] builds on investigations of adversarial AI as an attacking or defending strategy in cyber security systems, with a focus on defensive strategies countering attacked initiated intentionally through adversarial inputs with a deception aware AI; and exploring improved security frameworks relative to a cyber-adversary AI. The authors of [13] provide an extensive review of the current literature on AI-enabled cyber deception (CYDEC-AI) and identify publication gaps, specifically noting future possibilities to be applied at scale, and then propose conceptual integrated models embedding deception as a central informative dimension in autonomous cyber defense.

TABLE I. Summary of Literature Survey on AI in Cybersecurity

Ref	AI Techniques Used	Key Contributions
[1]	ML, DL, IDS, Neural Networks	Comprehensive review of AI-enabled threat detection, malware classification, and intrusion prevention
[2]	Statistical Drift Detection, ML Classifiers	Achieved 95.2% accuracy and 94% F1-score using adaptive malware detection in IIoT
[3]	CAS, TAM, TPB	Conceptual AI framework to promote threat awareness and resilience in IoT networks
[4]	MDATA, ACAM Framework	Enhanced detection of complex cyber attacks and reduced false positives with alarm correlation modules
[5]	ML classifiers, ensemble learning	Assesses AI deployment challenges in enterprise systems and integration with existing IT
[6]	Adversarial AI, deception modeling	Introduces dynamic AI deception and attacker behavior analysis
[7]	Game Theory, ML	Models attacker-defender dynamics and optimizes AI deception deployment strategies

Ref	AI Techniques Used	Key Contributions
[8]	Risk analysis, AI trust models	Discusses trust, transparency, and risks of adversarial AI in cybersecurity systems
[9]	Normative frameworks, cyber law	Calls for regulation of AI deception systems under international cybersecurity norms
[10]	ML classifiers	Establishes foundational role of AI in intrusion detection and threat classification

III. Proposed AI Based Cyber-Security System: A Deception as a Defense Approach

The Proposed AI based cyber-security system operates as a complex system comprised of six interconnected modules that synergistically facilitate intelligent cyber deception and cyber security (see Figure 2).

- The Data Accommodation Layer serves as the 'sensor' for the system. It collects real time data input, such as network traffic, system logs and user behavior as data, for the AI Threat Intelligence Engine to utilize.

- The AI Threat Intelligence Engine then utilizes deep learning and machine learning algorithms to identify anomalies, predict the intent of the attacker and iteratively learn new attack patterns from the wide array of data.

- At the core of deception is the Deception Management System. Through AI, the Deception Management System deploys and manipulates real deceptive resources such as honeypots, honeytokens and decoy networks that appear as organic functions of the actual system.

- The Deception Management System therefore supports the Behavioral Analytics & Monitoring module which collects data from the attacker interactions and establishes profiles of attackers in order to enhance detection performance.

- The Automated Response Module absorbed the information from deception (the adaptive smart response) and executes the cue for action or an automated response to the attack - the analogue actions for example could be; isolating a hampered node, alerting the Security Team, blocking the compromised traffic, etc.

- The final module, Feedback & Learning Loop, promotes a fresh impute of stored intelligence of the attacker behaviour (even if they were successful to gain access or stopped) into the operational theatre to allow the deception elements and AI intelligence in the framework to improve their strategies. Figure 2 presents a graphic representation of the proposed CDDS framework.

Figure 2 illustrates that there is a live connection between the real-time collection of data, AI-driven threat analysis, the placement of deception assets, and the autonomous response systems. The entire workflow begins with the Detection phase. In this phase, the AI engine identifies abnormal system behavior or indicators of near-term threats through predictive analytics. Once an anomalous behavior is identified, the deception deployment module is activated, which releases context-aware decoys, including honeypots or honeytokens that are consistent with the intended attacker's intent. Next, in the Engagement phase, the attacker engages with the deceptive decoys while unwittingly revealing their tactics. Engagement will be monitored during the Intelligence Extraction phase, during which the system and analysts will collect and analyze tools, tactics, and procedures (TTPs). Finally, the Autonomous Defense phase involves actions taken when intelligent analysis prompts the system to quarantine something, notify the security team, or modify a future defense decision reflecting the new findings. In this manner, the entire process is a feedback loop to provide enhanced intelligence into the system.

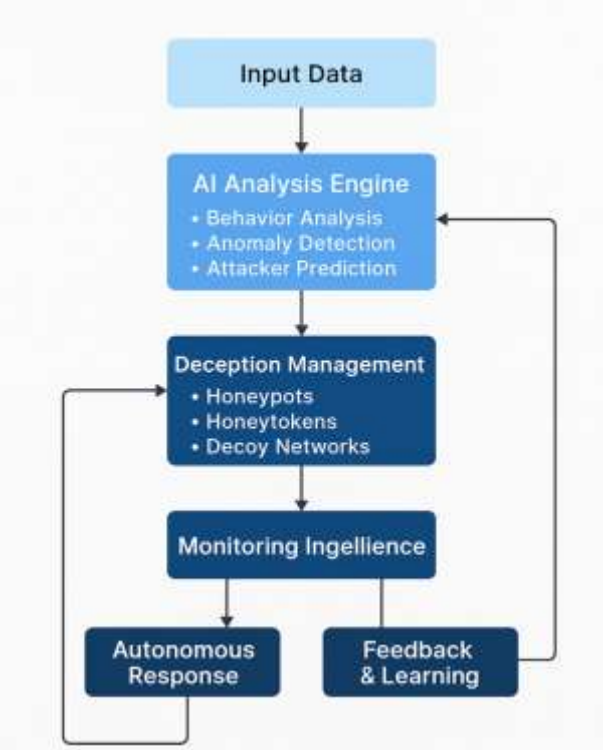

Figure 2: Proposed AI Based Cyber-Security System

A. *Mathematical Formulation*

This subsection contains the equations thatconnect the modules mathematically: anomaly detection triggers deception, deception generates intelligence, and intelligence updates the AI models in a continuous learning loop.

Let X_t represent the system state at time t (network traffic, logs, user actions, etc.). The AI engine estimates the likelihood of normal behavior:

$$P(X_t \mid \theta) = f_\theta(X_t) \quad \text{----------(1)}$$

An anomaly is detected if:

$$\text{Anomaly Score } A_t = 1 - P(X_t \mid \theta) > \tau \quad \text{----------(2)}$$

where τ is a predefined threshold.

Model the attacker's intent as a hidden state S_t in a probabilistic framework:

$$P(S_t \mid X_1{:}t) = \Sigma\, P(S_t \mid S_{t-1}) * P(S_{t-1} \mid X_1{:}t_1) \text{----------} (3)$$

This allows the AI to predict next steps of the attacker.

Let D_t denote the set of deployed deceptive resources at time t, and $E(D_t)$ their effectiveness in misleading an attacker:

$$E(D_t) = \alpha * \text{Interaction Rate} + \beta * \text{Time Engagement} + \gamma * \text{Data Extracted from Attacker} \text{ ----------} (4)$$

where α, β, γ are weighting factors to quantify impact.

The system updates its knowledge base K based on attacker interactions:

$$K_{t+1} = K_t + \eta * \Delta K_t \text{---------- ----------} (5)$$

where η is the learning rate, and ΔK_t represents newly extracted intelligence from attacker behaviour.

IV. Challenges and Ethical Considerations

There is a potential for artificial intelligence in systems that enable cyber deception, such as a cyber deception detection system (CDDS), to lead us to a new, or new-ish, defense paradigm, notwithstanding the implementation complexities that are both technical and ethical. One technical challenge we have to contend with is false positives. For example, our AI model may be based on anomaly detection templates; however, the AI model may misattribute typical user behaviors as malicious user behaviors, and result in triggering deception tools unnecessarily. False activations also have their consequences for example, confusion, fatigue in responding protocols, or degrading user trust. A second major technical challenge is scalability. We cannot simply scale a real-time, adaptive deception without a considerable computational footprint, and an algorithm that is adaptive and scalable rapidly when needed. This isn't sustainable for anyone and especially users that may be more limited by technology and fiscal constraints.

Although all forms of deception conclude with the goal of revealing something condemning, to authorize deception (be it a false system or false data), the ethics of potentially misleading understanding caused from misinterpretation, complications that stem from some level of tension, or limits of behavioral and/or legal expectations should be challenging. A second ethical consideration is whether the tension producing deception and the transparent organizational level creates anything for the success of deception. Expectations of accountability and oversight in democratic expectation are based on transparency. This is what creates the tension, the greatest support of successful deception.

Legitimate application of AI-enabled deception should at all times be constrained to international laws, ethical theories, and governance frameworks that are based on transparency, proportionality, and respect for digital rights. In the absence of these mechanisms, negative ramifications on perceived validity and legality in the activity of cybersecurity will manifest even for innocuous deception.

V. Conclusions

Cyberattacks are becoming more advanced, stealthy, diverse, and frequent, and we begin to shift from traditional defensive and reactive approaches to more proactive and intelligent security models. In this paper we have defined and described a CDDS as a new security model that positions deception at the forefront of the design of a cyber defense system. CDDS includes the dynamic use of decoys; cognizance of adversarial thinking and behavior; independent, if necessary, reaction; and in effect, independent management of defense throughout. As explained throughout, CDDS provides layered defense, which counters intrusions, but also turns encounters into increased intelligence reassurance, in effect building resilience over time. If CDDS incorporates aspects of artificial intelligence, including, but not limited to, predictive modeling, adaptive learning and real-time decision making, then it would slow, confuse, and ultimately outthink the adversary. Regardless, the potential of CDDS to disrupt the way we defend cyber attacks is large. If these solutions are built ethically and legally right and used properly ultimately "deceive to defend" can be a new progression in the ongoing struggle to secure cyberspace against increasingly capable threats.

References

[1] S. Zeadally, E. Adi, Z. Baig, and I. A. Khan, "Harnessing artificial intelligence capabilities to improve cybersecurity," *IEEE Access*, vol. 8, pp. 23817–23840, 2020.

[2] A. E. Molina, "AI in Cybersecurity: Revolutionizing Threat Detection and Response," *Cloud Security Alliance*, Mar. 14, 2025.

[3] S. Kalsha, "AI in Cybersecurity: Smarter Threat Detection in 2025," Qodequay, Jul. 24, 2025.

[4] A. Tellache et al., "Advancing Autonomous Incident Response: Leveraging LLMs and Cyber Threat Intelligence," arXiv, Aug. 14, 2025.

[5] S. H. Ahmed, et al., "Statistical drift detection for identifying malware in industrial IoT," *Future Gener. Comput. Syst.*, vol. 123, pp. 17–28, 2021.

[6] A. Ullah, M. Younus, S. A. Madani, and W. Ahmed, "AI-driven threat detection in smart cities: A conceptual framework," *Comput. Electr. Eng.*, vol. 104, 2023.

[7] Y. Li, et al., "Artificial intelligence-enabled cybersecurity defense for smart cities: A novel attack detection framework based on MDATA," *J. Cybersecurity Privacy*, vol. 2, no. 1, pp. 59–78, 2022.

[8] R. Sahu and M. Qamar, "Securing the digital world: AI-enabled malware and intrusion detection for smart infrastructures," *J. Netw. Comput. Appl.*, vol. 189, 2023.

[9] D. L. Antunes and S. L. Sanchez, "The age of fighting machines: The use of cyber deception for adversarial artificial intelligence in cyber defence," *Front. Artif. Intell.*, vol. 6, 2023.

[10] M. Bonfanti, M. Dunn-Cavelty, and A. Wenger, "Artificial intelligence and cyber-security: A complex relationship," *Contemp. Security Policy*, vol. 42, no. 1, pp. 6–28, 2021.

[11] A. Ebunoluwa and A. James, "AI-Powered Honeypots: Enhancing Deception Technologies for Cyber Defense," 2025. [Online].

[12] S. A. Syed, "Adversarial AI and cybersecurity: Defending against AI-powered cyber threats," 2025.

[13] P. B. López, P. Nespoli, and M. G. Pérez, "Cyber Deception powered by Artificial Intelligence: Overview, Gaps, and Opportunities," 2024.