World Journal of Advanced Engineering Technology and Sciences
eISSN: 2582-8266
Cross Ref DOI: 10.30574/wjaets
Journal homepage: https://wjaets.com/

(REVIEW ARTICLE)

AI-driven security architecture: Innovations in autonomous threat response

Gowtham Kukkadapu *

InfoGravity, USA.

World Journal of Advanced Engineering Technology and Sciences, 2025, 15(03), 1798-1806

Publication history: Received on 06 May 2025; revised on 14 June 2025; accepted on 16 June 2025

Article DOI: https://doi.org/10.30574/wjaets.2025.15.3.1090

Abstract

This article presents a comprehensive overview of AI-driven security architectures and innovations in autonomous threat response. As cybersecurity landscapes evolve with increasingly sophisticated threats, traditional security approaches relying on signature-based detection and human intervention prove inadequate against modern attack methodologies. The paradigm shift toward autonomous security systems leverages machine learning and artificial intelligence to enable continuous adaptation and proactive defense mechanisms. The article examines foundational components of AI-driven security architectures, key innovations including reinforcement learning, generative adversarial networks, security orchestration platforms, and implementation strategies and best practices. While highlighting transformative potential, the article also addresses significant challenges, including model interpretability, adversarial vulnerabilities, computational constraints, and ethical considerations that security practitioners must navigate when deploying these advanced systems.

Keywords: Autonomous threat response; Machine learning security; Generative adversarial networks; Security orchestration; Adversarial resilience

1. Introduction

The cybersecurity landscape is experiencing unprecedented transformation due to the rapid evolution of sophisticated threats. Organizations worldwide face increasingly complex challenges as cybercrime rises, with global damages projected to reach unprecedented levels by 2025 [1]. Traditional security approaches that rely on human intervention and rule-based systems are proving inadequate against modern cyber-attacks, which are growing in sophistication and frequency. Studies indicate that security teams must process massive alerts daily, with many legitimate threats going undetected due to alert fatigue and analysis limitations [1].

AI-driven security architectures represent a paradigm shift in this domain, leveraging machine learning (ML) and artificial intelligence (AI) to enable systems that can autonomously detect, analyze, and respond to emerging threats with minimal human oversight. These systems demonstrate significant advantages in processing security telemetry at scale, identifying subtle patterns that human analysts might miss, and reducing response times during critical security incidents [2]. Integrating advanced analytics with autonomous response capabilities allows organizations to establish more proactive security postures, essential for countering modern threat actors who deploy increasingly sophisticated methodologies.

The evolution from reactive to proactive security has become necessary as adversaries develop techniques to evade traditional detection methods. Recent research highlights how threat actors increasingly utilize automation themselves, creating attack variations that can bypass conventional security controls [1]. Meanwhile, the average time to identify and contain data breaches remains concerning, with significant financial implications for affected organizations across

* Corresponding author: Gowtham Kukkadapu

all sectors [2]. These extended exposure periods highlight the limitations of conventional approaches and underscore the need for more responsive security architectures.

This article examines recent innovations in autonomous threat response systems, including reinforcement learning for adaptive defense, generative adversarial networks for threat simulation, and security orchestration platforms for streamlined incident management. Implementing these technologies has demonstrated measurable improvements in threat detection accuracy, reduction of false positives, and acceleration of incident resolution timeframes across multiple industry sectors [2]. Additionally, the article evaluates implementation challenges, including algorithmic transparency concerns, model maintenance requirements, and computational resource constraints that organizations must address when adopting AI-driven security architectures.

This article analyzes developments in autonomous security systems and provides insights for researchers advancing theoretical foundations and practitioners seeking to transform security operations through AI-driven architectures.

2. Foundations of AI-Driven Security Architecture

2.1. Evolution from Traditional Security Models

Traditional security architectures have historically relied on signature-based detection, predefined rules, and human analysis for threat mitigation. These approaches, while foundational, suffer from inherent limitations in modern threat landscapes where attacks evolve rapidly. Signature-based detection systems struggle with new malware variants, as they can only identify threats that match known patterns, while human-centered analysis cannot scale to process the volume of alerts generated in enterprise environments [3]. Rule-based systems require continuous manual updates, creating significant operational overhead for security teams who must constantly adjust detection rules rather than focusing on investigating actual security incidents.

AI-driven architectures represent an evolutionary step that addresses these limitations through continuous learning and adaptation. Machine learning models can identify anomalous behaviors without predefined signatures, processing events at scales impossible for human analysts [4]. Deep learning implementations have shown promising results in reducing false positive rates compared to conventional rule-based systems, while anomaly detection algorithms can identify previously unknown attack patterns through behavioral analysis rather than signature matching [4]. These capabilities enable security operations to transition from reactive postures to proactive stances where systems autonomously detect and mitigate emerging threats before significant damage occurs.

2.2. Core Components of Autonomous Security Systems

Autonomous security systems typically comprise several interconnected components that function as an integrated ecosystem. Data collection mechanisms serve as the sensory foundation, gathering security telemetry across network flows, endpoint behaviors, authentication events, and application logs [3]. This raw data undergoes preprocessing to standardize formats and extract relevant features while preserving critical security indicators.

ML processing engines represent the analytical core of autonomous systems, employing various algorithmic approaches to detect anomalies and classify threats. Research has shown that ensemble methods combining multiple algorithms often achieve higher accuracy than single-algorithm approaches [3]. Supervised models excel at classifying known attack patterns, while unsupervised techniques help identify previously unknown threats. Implementing these systems requires significant computational resources to maintain real-time analysis capabilities across enterprise networks.

Decision-making frameworks translate analytical findings into actionable security responses, employing confidence scoring mechanisms consider multiple contextual factors when evaluating potential threats [4]. These frameworks incorporate risk quantification models that assign values to potential security incidents based on asset criticality, threat severity, and exploitation probability. Modern implementations leverage statistical methods to calculate dynamic risk scores that evolve as new information becomes available.

Response orchestration platforms execute mitigation actions based on framework decisions, implementing response actions across network, endpoint, and identity domains without human intervention. These platforms can significantly reduce mean-time-to-respond by eliminating procedural delays and human decision latency [4].

2.3. Theoretical Frameworks for Autonomous Decision-Making

The theoretical underpinnings of autonomous security systems draw from various disciplines, creating a multifaceted foundation for operational capabilities. Decision theory provides mathematical models for evaluating uncertain outcomes, with security implementations using utility calculations to determine optimal responses [3]. These decision models can process decisions rapidly, enabling responses at machine speed rather than human timescales.

Game theory frameworks model adversarial interactions, treating security as a continuous competition between defenders and attackers. These models enable defensive systems to allocate resources optimally across potential attack surfaces, increasing protection for critical assets without requiring additional security infrastructure [3].

Cognitive science principles inform how autonomous systems process and contextualize security information. Security implementations incorporate architectures that model attention, memory, and reasoning processes, enabling systems to prioritize significant security events more effectively than traditional severity-based triage [4].

Machine learning theory provides the algorithmic foundation for autonomous systems, with implementations leveraging both classical approaches and deep learning architectures. Neural network models have shown promising results in distinguishing benign from malicious traffic patterns [4]. Reinforcement learning frameworks demonstrate particular promise, with research showing improvement in response effectiveness through continuous adaptation to evolving attack methodologies.

Table 1 AI-Driven vs. Traditional Security Performance [3,4]

Security Metric	Performance Value
Traditional Detection Accuracy	Low
AI-Driven Detection Accuracy	High
Traditional Response Time	Slow
AI-Driven Response Time	Fast
False Positive Rate Reduction with AI	Significant

3. Key Innovations in Autonomous Threat Response

3.1. Reinforcement Learning for Adaptive Defense

Reinforcement learning (RL) has emerged as a powerful approach for developing adaptive defense mechanisms in cybersecurity. Unlike supervised learning methods that require extensive labeled datasets, RL enables systems to learn optimal security policies through continuous interaction with dynamic environments [5]. This paradigm employs reward-based mechanisms, where defensive systems learn to maximize security objectives while minimizing resource utilization and operational disruption. RL's adaptability makes it particularly valuable in contexts where threat landscapes evolve rapidly and adversaries actively modify their techniques to evade detection.

Security systems employing RL demonstrate significant advantages in operational efficiency and threat mitigation. These systems can dynamically adjust firewall configurations, intrusion detection thresholds, and access control parameters in response to shifting threat indicators without requiring explicit reprogramming by security teams [5]. The autonomous nature of these systems enables them to learn from each encounter with malicious activity, continuously refining their defensive strategies through experience rather than manual configuration updates.

Recent advancements in deep reinforcement learning have expanded these capabilities by incorporating neural networks that can process complex, high-dimensional security data. These implementations have shown effectiveness against advanced persistent threats (APTs), which typically employ sophisticated evasion techniques to circumvent conventional security controls [6]. This improvement addresses one of the most persistent challenges in cybersecurity operations—balancing detection sensitivity against alert fatigue among security analysts.

3.2. Generative Adversarial Networks for Threat Simulation

Generative adversarial networks (GANs) represent another significant innovation in autonomous security, providing sophisticated threat simulation and defensive training mechanisms. GANs comprise two neural networks—a generator and a discriminator—that operate in a competitive framework to improve each other's capabilities through continuous iteration [5]. In cybersecurity applications, GANs excel at creating realistic simulations of attack scenarios that would be difficult or impossible to obtain through conventional means.

Security teams leverage GAN-generated synthetic malware samples, phishing campaigns, and network intrusion patterns to train defensive systems against potential zero-day exploits before they appear in attack scenarios. This proactive approach addresses a fundamental limitation of traditional security systems, which typically rely on historical attack data and struggle to identify novel threats [6]. By training on synthetically generated attack patterns that represent potential future threats, defensive systems can develop detection capabilities for attack methodologies they have never encountered in operational environments.

Research implementations of GAN-based training for intrusion detection systems (IDS) have demonstrated improvements in identifying novel attack vectors compared to systems trained exclusively on historical data. These enhancements are particularly pronounced when dealing with sophisticated evasion techniques that deliberately manipulate traffic patterns to avoid detection [5].

3.3. SOAR Platforms for Streamlined Incident Management

Security orchestration, automation, and response (SOAR) platforms represent a transformative approach to incident management in modern security operations. These platforms integrate AI-driven tools to automate complex incident response workflows that traditionally require extensive manual intervention [6]. SOAR implementations leverage machine learning algorithms to prioritize security alerts based on severity, impact, and organizational context, addressing the chronic challenge of alert overload that affects many security operations centers.

Modern SOAR platforms incorporate sophisticated capabilities for correlating disparate threat intelligence from multiple sources, creating comprehensive threat narratives that provide security analysts with contextual understanding. This correlation functionality employs machine learning techniques, including natural language processing (NLP) for analyzing unstructured threat data from security blogs, forums, and research reports [5]. By synthesizing information across diverse sources, SOAR platforms can identify connections between seemingly unrelated security events, revealing coordinated campaigns that might otherwise remain undetected.

Recent advancements in SOAR include predictive analytics capabilities that anticipate attack progression patterns based on initial indicators, enabling preemptive intervention before attacks reach critical systems [6]. These predictive models analyze historical attack sequences and leverage current threat intelligence to forecast likely attack paths through organizational infrastructure. By identifying potential attack trajectories, security teams can implement targeted countermeasures that block adversary progression while minimizing disruption to legitimate business operations. This capability transforms security operations from primarily reactive to a proactive discipline that anticipates and counters threats before they cause significant damage.

Table 2 Impact of AI Innovations on Cybersecurity Response [5,6]

Innovation Type	Primary Benefit
Reinforcement Learning	Adaptive defense without manual reprogramming
Deep Reinforcement Learning	Effective against advanced persistent threats
Generative Adversarial Networks	Realistic threat simulation for zero-day detection
GAN-based Training	Improved detection of novel attack vectors
SOAR Platforms	Automated incident response workflows

4. Implementation Strategies and Best Practices

4.1. Integration with Existing Security Infrastructure

Successful deployment of AI-driven security architectures requires thoughtful integration with existing security infrastructure to ensure operational continuity while enhancing defensive capabilities. Organizations must carefully evaluate compatibility with legacy systems that utilize different data formats, communication protocols, and operational paradigms [7]. This integration challenge extends beyond technical considerations to encompass organizational processes, security governance frameworks, and compliance requirements that constrain implementation options.

A phased implementation strategy has emerged as the most effective approach for introducing AI-driven security capabilities into established environments. This methodology begins with targeted deployments focused on specific use cases demonstrating clear value while minimizing disruption to critical security functions [7]. Initial deployments often focus on areas with well-defined security processes and abundant historical data, such as email security, endpoint protection, or network traffic analysis. As these initial deployments mature, organizations can progressively expand AI capabilities to address more complex security domains.

Integration challenges frequently center on data access and normalization requirements, as AI systems depend on comprehensive security telemetry to identify patterns and anomalies effectively. Organizations must establish secure data pipelines that gather relevant information from diverse security tools without creating new attack surfaces or compliance risks [8]. These pipelines must address data format inconsistencies, temporal alignment challenges, and access control requirements while maintaining performance under high-volume conditions.

4.2. Training Requirements and Data Considerations

The efficacy of AI-driven security systems depends heavily on the quality, quantity, and diversity of training data used during model development and refinement. Organizations must establish robust data management practices addressing technical and governance requirements to ensure model accuracy and reliability in operational environments [7]. These practices must account for the dynamic nature of security threats, where adversary techniques evolve continuously, potentially rendering historical training data less relevant.

Data collection strategies must balance breadth and depth, gathering diverse security telemetry while maintaining sufficient context for meaningful analysis. Effective implementations typically combine multiple data sources, including network traffic, endpoint events, authentication logs, and threat intelligence feeds to provide comprehensive visibility [7]. This multi-source approach enables detection of sophisticated attacks that manifest across different aspects of the security environment rather than appearing in isolation.

Labeling processes represent a challenge in security contexts, as accurately identifying malicious activities within historical data requires specialized expertise and significant time investment [8]. Organizations have adopted various approaches to address this challenge, including leveraging existing security tools to provide initial classifications and utilizing semi-supervised learning techniques that reduce labeling requirements.

Continuous retraining protocols are essential to maintain system performance as threat landscapes evolve and organizational environments change over time. Without regular updates, AI models experience performance degradation as new attack techniques emerge and legitimate usage patterns shift, a phenomenon known as concept drift [8]. Organizations establish systematic retraining schedules based on temporal factors and performance monitoring, triggering model updates when detection accuracy falls below established thresholds.

4.3. Performance Metrics and Evaluation Frameworks

Measuring the effectiveness of autonomous security systems requires comprehensive evaluation frameworks that assess multiple performance dimensions under various operational conditions. Organizations must develop metrics that align with security objectives while providing actionable insights for continuous improvement [7]. These frameworks should extend beyond traditional classification metrics to evaluate operational factors such as resource utilization, investigation efficiency, and resilience against evasion attempts.

Detection accuracy represents a fundamental performance dimension, typically measured through precision and recall metrics [7]. These measurements must account for class imbalance in security contexts, where legitimate activities typically outnumber malicious events by significant margins. Organizations often establish different performance

thresholds for various threat categories based on potential impact, accepting higher false positive rates for critical threats while requiring greater precision for lower-severity issues.

Response time metrics assess the operational effectiveness of autonomous systems, measuring intervals between initial threat indicators and defensive actions [8]. These measurements typically span multiple phases: initial detection, alert triage, contextual analysis, decision-making, and response execution. Organizations establish baseline expectations for each phase based on threat severity and complexity.

Resilience against adversarial attacks has emerged as a critical performance dimension for AI-driven security systems, as sophisticated adversaries increasingly employ evasion techniques specifically designed to manipulate machine learning models [8]. Evaluation frameworks must assess how systems perform when confronted with deliberately crafted inputs intended to trigger false negatives or false positives. Regular adversarial testing using techniques such as gradient-based attacks, input perturbations, and model poisoning attempts helps identify and address potential vulnerabilities before they can be exploited in real-world scenarios.

Table 3 Key Implementation Considerations for AI Security [7,8]

Implementation Aspect	Best Practice Approach
Integration Strategy	Phased deployment with targeted use cases
Data Management	Multi-source telemetry with context preservation
Model Maintenance	Continuous retraining to prevent concept drift
Performance Measurement	Balanced precision and recall metrics
Resilience Testing	Regular adversarial attack simulations

5. Challenges and Limitations

5.1. Model Interpretability and Explainability

A significant challenge in AI-driven security architectures is the "black box" nature of many advanced machine learning models, particularly deep learning implementations that utilize complex neural network architectures. Security practitioners often struggle to understand why systems make specific decisions, complicating trust, accountability, and regulatory compliance in high-stakes security environments [9]. This opacity becomes particularly problematic when autonomous systems generate false positives or miss sophisticated attacks, as security teams cannot easily determine whether these errors represent fundamental model limitations or temporary anomalies.

Research in explainable AI (XAI) aims to address these limitations by developing methods that provide human-understandable justifications for AI decisions without compromising system performance [9]. These approaches span multiple dimensions, including post-hoc explanation techniques that illuminate black-box models after training and inherently interpretable architectures that maintain transparency throughout the decision process. Current XAI implementations for security applications utilize various techniques, including feature importance rankings, identifying which aspects of security data most influence specific decisions.

Despite progress in XAI research, significant challenges remain in applying these techniques to complex security contexts where adversaries actively attempt to evade detection [10]. Explanations must balance comprehensiveness with cognitive accessibility, providing sufficient detail for security analysts to understand system reasoning while avoiding overwhelming complexity that obscures critical insights. Additionally, explanation mechanisms may create new attack surfaces, as adversaries could leverage explanation outputs to reverse-engineer detection models.

5.2. Adversarial Attacks and Model Vulnerabilities

Paradoxically, AI-driven security systems present attractive targets for adversaries, creating a meta-security challenge wherein defensive technologies require their protection. Attackers can exploit vulnerabilities in machine learning models through various techniques, creating an adversarial arms race between defensive innovations and evasion methods [9]. This dynamic highlights the importance of considering adversarial resilience as a fundamental design parameter for autonomous security systems rather than treating it as an optional enhancement.

Data poisoning attacks represent one common vulnerability vector. In these attacks, adversaries manipulate training data to introduce systematic biases or specific blind spots into security models [9]. These attacks can occur during initial model development if training data contains compromised examples or during operational retraining phases, where models incorporate new observations that may include adversary-manipulated inputs.

Evasion attacks present another significant challenge, as adversaries modify their attack patterns to avoid triggering detection by autonomous security systems [10]. For example, subtle modifications to network traffic patterns can cause ML-based intrusion detection systems to misclassify malicious activities as legitimate. These evasion techniques become increasingly sophisticated as adversaries gain understanding of underlying detection mechanisms, either through public research or through empirical testing against target systems.

5.3. Computational Overhead and Resource Constraints

The computational requirements of sophisticated AI models present practical challenges for many organizations implementing autonomous security architectures. Resource-intensive algorithms may strain infrastructure, particularly in real-time detection scenarios that demand low-latency responses to emerging threats [9]. This computational overhead becomes especially problematic at enterprise scale, where security systems must process massive data volumes across distributed environments with varying connectivity and processing capabilities.

Real-time detection requirements exacerbate computational challenges, as many security scenarios demand immediate analysis to enable timely intervention before attacks succeed [9]. The tension between analytical depth and response speed creates difficult tradeoffs, particularly for sophisticated threats requiring complex analysis to identify confidently. Organizations typically establish tiered detection architectures that employ lightweight models for initial screening, with progressively more sophisticated analysis for suspicious activities that warrant deeper inspection.

Edge deployment scenarios present particular challenges for AI-driven security, as endpoint devices and network appliances often have limited processing capabilities compared to centralized infrastructure [10]. Organizations increasingly require distributed security intelligence that can function effectively despite these constraints, protecting remote offices, mobile workforces, and environments that cannot rely exclusively on cloud-based security analysis.

5.4. Ethical and Regulatory Considerations

The autonomous nature of AI-driven security systems raises important ethical questions regarding accountability and oversight in security operations. When systems make autonomous decisions that impact security postures, determining responsibility for adverse outcomes becomes complex, creating potential accountability gaps that complicate incident management and liability determination [9]. These ethical concerns extend beyond technical considerations to encompass organizational governance, professional responsibility, and societal expectations regarding autonomous systems.

Responsibility allocation represents a fundamental ethical challenge in autonomous security, requiring clear delineation of human and machine roles across various operational scenarios [9]. Organizations must establish explicit policies regarding which decisions systems can make autonomously versus which require human approval, with these boundaries typically reflecting both technical capabilities and risk tolerance.

Privacy implications present another ethical dimension of autonomous security, as AI systems typically require extensive data access to function effectively [10]. This data often includes sensitive information about user behaviors, organizational operations, and communication patterns that create privacy risks if misused or inadequately protected. Organizations must carefully balance security requirements against privacy considerations, implementing data minimization practices, anonymization techniques, and access controls.

Emerging regulations regarding AI usage and data protection introduce compliance challenges that security architects must address proactively rather than reactively [10]. These regulatory frameworks increasingly establish specific requirements for autonomous systems, including transparency obligations, limitations on automated decision-making, data protection mandates, and audit requirements that affect security implementations.

Table 4 Major Challenges in AI Security Implementation [9,10]

Challenge Category	Primary Concern
Model Interpretability	"Black box" nature complicates trust and accountability
Adversarial Attacks	AI systems themselves become targets for sophisticated attackers
Data Poisoning	Manipulation of training data creates security blind spots
Computational Requirements	Resource-intensive algorithms strain infrastructure
Ethical Considerations	Autonomous decisions raise accountability questions

6. Conclusion

AI-driven security architectures represent a paradigm shift in cybersecurity, transforming reactive defense postures into proactive, autonomous protection systems. Integrating reinforcement learning, generative adversarial networks, and security orchestration platforms enhances threat detection capabilities, reduces response times, and improves security resilience against sophisticated threats. However, the transition to autonomous security systems involves navigating substantial challenges related to model interpretability, adversarial vulnerabilities, computational requirements, and ethical considerations. Organizations must balance the promise of AI-driven automation with appropriate human oversight and governance frameworks—the future direction points toward integration with zero-trust frameworks and adopting privacy-preserving techniques like federated learning. Interdisciplinary collaboration between AI experts, cybersecurity specialists, and ethicists will be essential to realize the full potential of autonomous threat response systems while mitigating associated risks, ultimately providing robust protection for increasingly interconnected digital ecosystems.

References

[1] Calif Sausalito, "2022 Cybersecurity Almanac: 100 Facts, Figures, Predictions And Statistics," Cybercrime Magazine, 2022. [Online]. Available: https://cybersecurityventures.com/cybersecurity-almanac-2022/

[2] Chris Brown, "The Real Cost of Data Breach in 2025," Viking Cloud, 2025. [Online]. Available: https://www.vikingcloud.com/blog/the-real-cost-of-data-breach

[3] Keshav Malik, "AI in Cybersecurity: Advantages, Limitations and the Future," Astra, 2025. [Online]. Available: https://www.getastra.com/blog/ai-security/ai-in-cybersecurity/#:~:text=Challenges%20and%20Risks%20of%20AI%20in%20Cybersecurity&text=The%20concept%20of%20Adversarial%20AI,targeted%20and%20more%20effective%20attacks.

[4] Enoch Oluwademilade Sodiya et al., "A comprehensive review of machine learning's role in enhancing network security and threat detection," World Journal of Advanced Research and Reviews 21(2), 2024. [Online]. Available: https://www.researchgate.net/publication/378288472_A_comprehensive_review_of_machine_learning's_role_in_enhancing_network_security_and_threat_detection

[5] Naveen Kumar Thawait, "Machine Learning in Cybersecurity: Applications, Challenges and Future Directions," International Journal of Scientific Research in Computer Science Engineering and Information Technology 10(3):16-27, 2024. [Online]. Available: https://www.researchgate.net/publication/380327525_Machine_Learning_in_Cybersecurity_Applications_Challenges_and_Future_Directions

[6] Jason Miller, "The Role of AI in Modern Cybersecurity," BitLyft Cybersecurity, 2024. [Online]. Available: https://www.bitlyft.com/resources/the-role-of-ai-in-modern-cybersecurity#:~:text=AI%2Ddriven%20predictive%20analytics%20help,to%20implement%20preventive%20measures%20proactively.

[7] Fortinet, "AI in Cybersecurity: Key Benefits, Defense Strategies, & Future Trends," Fortinet.com. [Online]. Available: https://www.fortinet.com/resources/cyberglossary/artificial-intelligence-in-cybersecurity

[8] Tahir, "Adversarial Machine Learning: Understanding Risks and Defenses," Medium, 2025. [Online]. Available: https://medium.com/@tahirbalarabe2/%EF%B8%8Fadversarial-machine-learning-a-taxonomy-and-terminology-nist-ai-100-2e2023-fb7ccc11ce98

[9] Javier Rando et al., "Adversarial ML Problems Are Getting Harder to Solve and to Evaluate," arXiv, 2025. [Online]. Available: https://arxiv.org/html/2502.02260v1

[10] HiddenLayer, "Understanding the Threat Landscape for AI-Based Systems," HiddenLayer.com, 2024. [Online]. Available:https://hiddenlayer.com/innovation-hub/understanding-the-threat-landscape-for-ai-based-systems/