# Abstract

a

# Introduction

In an age defined by the relentless increase of technology, the growing digital landscape has become both a playground for innovation and a battleground for cyber threats. As organizations increasingly rely on technology to operate, communicate, and store critical data, safeguarding these assets against potential adversaries becomes paramount. This realization has given rise to Penetration Testing (PT) as a vital and proactive strategy that allows organizations to simulate cyberattacks on their systems to discover and eliminate dangerous vulnerabilities. PT, often referred to as ethical hacking, is the “offensive approach” of probing and assessing computer systems, networks, and applications “to actively identify vulnerabilities and then exploit them in the same way as a genuine attacker [1].”

This situation has prompted a growing trend toward integrating automation technologies, including Artificial Intelligence (AI), Machine Learning (ML), and Reinforcement Learning (RL) into PT. Automated PT can significantly reduce the time and resources required for testing, making it a crucial development in the field. While extensive research has already begun to explore this integration [1] [3], many of these approaches still require manual human intervention for vulnerability identification. However, recent innovative methods have emerged that leverage neural networks to gain a deeper understanding of the intricate and dynamic security environments within modern networks, ultimately enhancing the role of AI in PT. This evolution is a pivotal moment for cybersecurity and AI as they converge to tackle digital threats.

# Related work (literature review)

# Proposed approach

intro

## *Project Scope*

The rapidly evolving cybersecurity landscape, with its complex and labor-intensive strategies, stands poised to significantly benefit from these recent advancements in AI technology. This surge in AI's significance is crucial for cybersecurity endeavors such as PT, the linchpin of proactive cybersecurity. The integration of AI in PT has the potential to level the playing field and empower defenders to anticipate and mitigate threats effectively.

The motivation behind this project is to comprehensively explore modern threats and current PT methods by examining the integration of various AI techniques including ML, RL, and Deep Reinforcement Learning (DRL), into PT practices. At its core, this project aims to address a fundamental question: How can AI techniques be effectively harnessed throughout the PT process? This question is critical, especially with the increasing adoption of AI by cybercriminals, as it demands cybersecurity professionals to adapt promptly; a sole reliance on traditional PT methods may leave pentesters unequipped to combat modern threats. This project aims to benefit cybersecurity defenders, ethical hackers, security analysts, and researchers by providing insights into AI tools and techniques, along with suggestions for overcoming current limitations.

Additionally, the hope is that a shift towards intelligent automation can not only reduce testing time and resources, but also mitigate many of the prevalent and "recurrent human errors" in manual PT that stem from factors such as "tiredness, omission, and pressure" [1]. This transformation in the field, as highlighted by [3], signifies the use of "advanced algorithms, machine learning, and AI to scan systems for vulnerabilities" and offering a path towards more effective, efficient, and error-resistant cybersecurity practices

## *Current Limitations*

However, addressing the current landscape of research in the integration of AI in PT reveals several notable limitations. Firstly, there exists a significant gap in research and testing within this domain, thereby signifying an unexplored terrain ripe for investigation. While research has begun exploring this novel integration [1] [3], many of these approaches are limited by their reliance on manual human intervention for vulnerability identification and exploitation. While this reliance hinders the full realization of AI's potential in the PT process, transitioning towards a fully autonomous solution presents complex technical hurdles. For example, the shift to intelligent, real-time detection would require addressing issues of data accuracy, model robustness, and environment adaptability.

To ensure data accuracy, it is essential that all input data is precise and reliable. However, achieving this precision in dynamic network environments can be challenging since data is often incomplete, outdated, or biased; this is especially dangerous in PT as these inaccuracies can lead to the misinterpretation of vulnerabilities or threats and potentially result in ineffective security measures. Therefore, data accuracy is vital for ensuring model robustness. For truly effective threat detection, models must be capable of adapting to evolving threats and environments in order to perform consistently across diverse scenarios. This level of adaptability requires sophisticated algorithms that can understand and interpret incomplete or biased data while accounting for various attack vectors, software vulnerabilities, and system configurations.

The integration of AI in PT is a relatively novel field, where these challenges of ensuring accuracy, robustness, and adaptability are still being navigated and tested. Due to its infancy, research in this domain is notably limited and requires further exploration and innovation before a fully automated PT tool becomes reality. As this report delves into the Literature Review, it will examine existing solutions to these discussed challenges. Then, by exploring current research in the field, this report can identify existing solutions and use their insights to pave the way for innovative approaches to overcome current limitations and advance the field of AI in PT.

# Implementation and evaluation

Intro (ADMS,

## *Automated Decision-Making Systems*

Automated Decision-Making Systems (ADMS) are foundational in the evolution of AI for penetration testing due to their ability to reduce human reliance while simultaneously enhancing the discussed critical limitations such as data accuracy, model robustness, and environmental adaptability. These systems leverage techniques such as deep learning, transfer learning, text-mining, NLP, and more to automate critical decision-making processes and allowing for faster and more accurate identification of potential vulnerabilities and weaknesses in cybersecurity defenses [ram].

This automation is important for pentesters as it significantly minimizes manual intervention, thereby reducing the potential for errors and improving the overall effectiveness of AI models in penetration testing scenarios. For example, ADMS can leverage formal decision support systems (DSS) and attack graph modeling to assist pentesters in identifying optimal attack vectors and vulnerabilities [ram]. By automating the process of evaluating risk factors, integrating vulnerability intelligence extracted from public vulnerability datasets such as CVE and NVD, and determining the best course of action, ADMS streamline and enhance the accuracy of penetration testing strategies.

For example, recent work by Georgescu et al. demonstrates the effectiveness of automation in identifying vulnerabilities in Internet of Things (IoT) systems [george]. They use a named entity recognition (NER)-based solution to analyze security data sources in natural language, which can be time-consuming for security experts. The proposed system incorporates considerations for 'ongoing changes in the CVE database and the current situation of the IoT system,' to assist pentesters in quickly identifying potential vulnerabilities, assessing their impact, and recommending targeted security measures [ram].

Overall, ADMS play a crucial role in overcoming key challenges associated with fully automating penetration testing, including excessive human reliance, ensuring data accuracy, bolstering model robustness, and enhancing environmental adaptability. Through advanced AI techniques such as deep learning and natural language processing, ADMS streamlines decision-making processes and reduces human reliance, thereby minimizing errors and enhancing the overall effectiveness of AI models in penetration testing scenarios. This automation not only speeds up the identification of vulnerabilities but also ensures a more accurate and adaptable approach to cybersecurity testing, making ADMS indispensable for offensive pentesters in navigating the dynamic and complex cybersecurity landscape.

## *Self-Improving Systems*

The integration of self-improving systems is another important part of enhancing the effectiveness of AI within offensive pentesting methods by significantly reducing human input, enhancing data accuracy, increasing model robustness, and improving environmental adaptability. As highlighted by [real], there has been a recent inclination towards integrating ML models into feedback loops through methodologies such as reinforcement learning (RL) to interact with environments, take actions, receive feedback, and adapt accordingly. This approach allows AI applications to "operate in real environments", "react to sensory data", "perform continuous micro-simulations", aand execute contextually appropriate actions [real]. By incorporating this action-reward framework into automated systems, researchers provide a means to quickly detect and respond to threats - all with millisecond latency and high throughput, while minimizing human oversight. Therefore, these systems not only reduce overall manual workload, but they also minimizes the risk of human error, thereby ensuring data accuracy while augmenting model robustness and environmental adaptability.

As mentioned, these systems leverage real-time data to refine their decision-making processes continuously in order to adapt to changes and learn from interactions within their environment. This capability is crucial in penetration testing where dynamic assessment and response are paramount. As a result, these systems use their low latency, high throughput, and robust fault tolerance capabilities to efficiently handle important model requirements such as "dynamic task creation", "heterogeneous tasks", and "arbitrary dataflow dependencies [real]". These attributes are essential for the real-time analysis of security vulnerabilities, rapid threat identification, as well as managing the complex computational demands inherent to pentesting scenarios.

Self-improving systems also improve the robustness of AI models by enabling them to handle diverse and dynamic computational tasks. Task graphs, for instance, are utilized within real-time ML systems to orchestrate heterogeneous tasks, which vary in complexity, resource requirements, and processing types - all while meeting stringent performance benchmarks [real]. Essentially, this capability ensures that the AI systems can maintain high performance and reliability, even when faced with varying operational demands. Therefore, by introducing self-improvement mechanisms, AI models gain a unique flexibility to dynamically adjust their computational strategies. This not only increases the model's robustness by allowing it to maintain optimal performance across a variety of conditions, but also enhances its environmental adaptability.

## *Human-in-the-loop*

In addition to Automated Decision-Making Systems and Self-Improving Systems, Human-in-the-loop (HITL) systems offer an alternative approach to overcoming the challenge of overt human reliance in offensive pentesting methods. While the former systems prioritize reducing human input through automation and self-improvement, HITL systems acknowledge the value of human expertise and decision-making capabilities. By "integrating human knowledge and experience" into AI models, HITL systems adopt a hybrid approach that not only enhances data accuracy, model robustness, and environmental adaptability, but also ensures that AI-driven penetration testing tools are guided by human expertise, rather than being weighed down by it.

HITL systems are particularly successful at refining the capabilities of AI models in fields requiring high accuracy and adaptability, such as pentesting. By incorporating human intelligence into the ML loop during the training phase, the model is able to improve understanding and performance. As highlighted by Wu et al. [wu], human intervention during this phase is vital for the success of deep learning, given its significant reliance on high-quality data and precise parameter adjustments. This is especially true when dealing with complex, high-demensional data where finding optimal parameter can be challenging and often requires significant expertise. However, HITL can address this challenge through an "interactive parameter adjustment mode," which allows users to actively participate in refining parameters based on their intentions or goals.

The reason HITL systems are a viable option for mitigating challenges caused by excessive human input is that, despite initially seeming counterintuitive due to an emphasis on human involvement, these systems capitalize on AI models trained within them becoming adept at mimicking human decision-making. In other words, as AI models trained under these systems become more proficient at simulating human-like decision-making processes, the need for continuous human intervention gradually decreases. This is especially crucial for security systems, which rely heavily on human operation, as human errors are often amplified due to factors such as "inattentiveness" or "nonproficiency" [wu]. Therefore, through the collaborative learning process with humans, HITL systems improve their ability to autonomously predict and manage operational states. For instance, wu et al. [wu] illustrate how automated feedback loops that interpret control panel interactions provide a real-time, precise monitoring and response mechanism that an AI trained without human-derived insights might miss. This precision is vital for pentesting because it allows for the immediate detection and mitigation of security vulnerabilities that could be exploited in real attacks.

HITL also significantly increases the robustness of AI models. As these systems encounter a diverse array of scenarios, they learn, adapt, and strengthen their predictive and reactive capabilities. This adaptability is vital in offensive penetration methods, where AI systems must robustly handle unpredictable and evolving security threats. By continuously learning and adapting from human experts within the HITL framework, these systems develop resilience and flexibility. By dynamically allocating roles and responsibilities between humans and AI—effectively deciding "who should do what" based on situational demands—these systems can adapt to changing environmental conditions and requirements [Wu]. In offensive penetration, such adaptability translates to AI systems capable of navigating and securing diverse network environments and configurations, which are subject to constant change due to technological advancements and evolving threat landscapes.

Therefore, by integrating HITL, AI systems achieve higher data accuracy, increased robustness against diverse threats, and improved adaptability to different operational environments. These enhancements are crucial for the effectiveness of AI in offensive penetration methods, enabling these systems to preemptively identify vulnerabilities, adapt to new attack vectors swiftly, and respond with precision. Ultimately, this integration leads to reduced reliance on continuous human input and provides more secure and resilient network infrastructures.

# Comparison

a

# Conclusion and discussions

a

[ram] <https://www.sciencedirect.com/science/article/pii/S1566253523001136#sec0005>

[George] <https://www.mdpi.com/1424-8220/19/15/3380>

[real] Real-Time Machine Learning: The Missing Pieces