# Abstract/ Overview

* Systems
  + Automated Decision-Making,
  + Self Improving Systems,
  + Human-in-the-loop
* discuss
  + Comprehensive Assessments:
    - comprehensive assessment of modern threats/current PT methods/new systems
      * ML, RL. DR
    - how AI techniques can be effectively harnessed throughout the PT process to conduct comprehensive assessments of cybersecurity defenses and vulnerabilities
    - systems
      * how ADMS and Self-Improving Systems contribute to comprehensive assessments in PT by automating critical decision-making processes, identifying vulnerabilities quickly, and adapting to evolving threats.
  + Operational Efficiency:
    - Motivation - streamlining operations/reducing errors associated with human factors.
    - how AI integration in PT can lead to intelligent automation > reducing testing time/resources, and mitigating prevalent human errors in manual PT.
    - ADMS –
      * reduces manual intervention, minimizes errors, and improves the overall effectiveness of AI models in PT scenarios
      * streamlining decision-making processes, evaluating risk factors, and recommending security measures to enhance operational efficiency
    - Self-Improving Systems
      * leveraging real-time data to continuously refine decision-making processes, adapt to changes, and learn from interactions within their environment (ie operational efficiency)
  + Vulnerability Management
    - !! human intervention, data accuracy, model robustness, and environmental adaptability
* Analysis
  + Empirical Analysis:
    - Explore modern threats, current PT methods, current/future systems
    - specific examples and projects that demonstrate the effectiveness of AI in identifying vulnerabilities and adapting to changing threats
      * Georgescu et al. => identifying vulnerabilities in Internet of Things (IoT) systems
  + Theoretical Inquiry:
    - theoretical foundations of AI integration in PT,
    - address critical challenges
    - how AI techniques can be harnessed throughout the PT process
    - potential benefits/challenges
    - need for further exploration and innovation in this domain
* Comparison Against Real-World PT Scenarios:

# 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)

A

# Proposed approach

A

## 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.

## Proposed Methodology and Solutions

### Section goal: Addressing limitations in previous section

### Reducing Human Input

* + Automated Decision-Making Systems
    - Intelligent Automation and Predictive Analytics: Develop AI systems capable of autonomously detecting vulnerabilities, assessing their severity, forecasting potential breaches through predictive analytics, and implementing preemptive mitigation strategies without human intervention.
    - <https://www.sciencedirect.com/science/article/pii/S1566253523001136>
      * “Predictive intelligence is intelligence that is actionable and relevant in a given context and can be used to anticipate attacks. Intrusion prediction tools are helpful in providing an active defence against future attacks by predicting the type, intensity and target of an intrusion in advance. Researchers are using deep-learning [73], [74], [75], [76] approaches to forecast the alerts from malicious sources [73] or on a given target [74], [75], [76] using the sequence of previous alerts [73], historic spam e-mail [74], and network traffic [75,76] data.”
      * “Malware prediction involves methods to predict and block the malicious files before they execute their payload completely, to prevent malware attacks rather than remedy them. In this direction, Rhode et al. developed a malware prediction model based on a recurrent neural networks (RNNs) model to predict malicious behaviour using machine activity data [77].”
      * “Attack prediction is deemed to have excellent potential for proactively advancing cyber resilience. Researchers present attack prediction schemes by utilizing different types of data retrieved from news sites and websites [78], dark web forums [79], national vulnerability databases [79], incident reports [79], and common vulnerabilities and exposure databases [80].”
    - *“*
    - *Automated Decision-Making Systems (ADMS) are instrumental in reducing reliance on human input and enhancing the effectiveness of AI models used for penetration testing. These systems leverage AI techniques to automate critical decision-making processes, allowing for faster, more accurate, and proactive responses to cybersecurity threats. One key area where ADMS proves invaluable is in risk planning, where formal decision support systems and attack graph modeling assist security planners in making economic comparisons with the cost of countermeasures and available risk budgets. For example, Rees et al. [81] utilized a genetic algorithm to find the optimal combination of countermeasures, enabling users to determine the preferred trade-off between investment cost and resulting risk.*
    - *In supply chain risk management, ADMS supports risk decisions by automating threat analysis, optimal cybersecurity investments, and assessments of cyber resilience. This comprehensive view of threats and vulnerabilities is essential for identifying, assessing, and managing supply chain risks effectively. Researchers, as highlighted by the research, actively use AI techniques for these tasks, including optimal cybersecurity investment strategies [86], [87], and assessments of cyber resilience [89].*
    - *Automated access control is another critical area where ADMS plays a pivotal role in reducing reliance on human input. By automating access control states, role mining, and situation-aware decision making, ADMS ensures that system access is restricted to authorized users based on their roles and regulations within an organization. For instance, researchers have proposed using AI techniques to update and maintain access control states [100], optimizing plans of actions to reconfigure access control based on exceptions or violations.*
    - *Cybersecurity awareness and training also benefit significantly from ADMS, as it enables adaptive and personalized training, awareness, and recommendations through natural language processing algorithms [104], [105]. This automation ensures that personnel and partners can carry out their information security duties and responsibilities in compliance with policies and procedures, reducing the need for manual oversight and intervention.*
    - *In conclusion, Automated Decision-Making Systems are crucial for reducing reliance on human input in penetration testing. By automating critical decision-making processes in risk planning, supply chain risk management, access control, awareness, and training, ADMS enhances the efficiency, accuracy, and scalability of AI-driven penetration testing tools, ultimately bolstering cybersecurity defenses and resilience against evolving threats.”*
  + Self-Improving Systems
    - Feedback Mechanisms for Optimization: Utilize machine learning techniques like online learning and continuous model training to refine algorithms based on new data continuously. Establish automated feedback mechanisms, including reinforcement learning, to learn from successes and failures and improve system performance iteratively.
    - **Real-Time Machine Learning: The Missing Pieces**
      * “ML applications are expanding from the supervised learning paradigm, in which static models are trained on offline data, to a broader paradigm, exemplified by reinforcement learning (RL), in which applications may operate in real environments, fuse and react to sensory data from numerous input streams, perform continuous micro-simulations, and close the loop by taking actions that affect the sensed environment.”
      * Using a “a logically-centralized control plane and a hybrid scheduler” to a flexible Dynamic task creation, Heterogeneous tasks, Arbitrary dataflow dependencies
      * system architecture that supports this programming model and meet our performance requirements (R1-R2) without giving up key practical requirements (R6-R7).
    - *“The proposed system architecture, with its focus on critical performance, execution model, and practical requirements for emerging ML applications like those used in penetration testing, plays a vital role in reducing reliance on human input. By achieving low latency, high throughput, and fault tolerance while handling dynamic task creation, heterogeneous tasks, and arbitrary dataflow dependencies, the architecture enhances the efficiency and effectiveness of AI models in penetration testing. These improvements are essential for real-time analysis of security vulnerabilities, rapid threat identification and mitigation, and managing complex computational tasks inherent in penetration testing scenarios. The architecture's ability to deliver millisecond-level performance and significant speedups leads to more accurate, responsive, and scalable AI-driven penetration testing tools, ultimately bolstering overall cybersecurity posture.*
    - *The system's approach to achieving real-time capabilities for ML applications is crucial in reducing human intervention in penetration testing tasks. The API and execution model support the creation of remotely executable tasks with dataflow dependencies, enabling non-blocking, asynchronous task execution without constraining throughput or computation graph dynamics. The inclusion of multiple worker processes, local schedulers, global schedulers, and an object store, all managed by a logically-centralized control plane, further reduces the need for manual oversight. Leveraging sharding in the database and a hybrid scheduling approach optimizes task assignment, leading to sub-millisecond scheduling latencies and efficient resource utilization critical for AI models used in penetration testing.*
    - *By reducing reliance on human input, the system enables AI models to autonomously handle real-time analysis, rapid response to security threats, and dynamic task management. This autonomy enhances scalability, adaptability, and agility in addressing evolving cybersecurity challenges, ensuring that AI-driven penetration testing tools can stay ahead of emerging threats and maintain robust security defenses. The system's detailed approach to efficient task execution and resource utilization is pivotal in minimizing manual intervention, streamlining operations, and improving overall cybersecurity posture in penetration testing environments.”*
  + Extra considerations
    - Continuous Improvement: Use audits, reviews, and Human-in-the-loop (HITL) Systems to ensure compliance with cybersecurity protocols and ethical standards during the gradual integration of automated systems. Start with less critical tasks to gauge effectiveness and fine-tune operations based on real-world feedback. Continuously update AI systems with new data and scenarios while training staff to use and trust automation tools effectively, understanding when and how to intervene in automated processes.
    - <https://www.sciencedirect.com/science/article/pii/S0167739X22001790#sec1>
      * “Humanin-the-loop aims to train an accurate prediction model with minimum cost by integrating human knowledge and experience. Humans can provide training data for machine learning applications and directly accomplish tasks that are hard for computers in the pipeline with the help of machine-based
      * approaches.”
    - While full automation is still a ways away, we could implement HITL and gradually transition to a full and fine-tune operations based on real-world feedback.
    - *Human-in-the-loop (HITL) systems present a sophisticated blend of human expertise and automated processes, pivotal in refining the capabilities of AI models, particularly in fields requiring high accuracy and adaptability, such as penetration testing. HITL systems incorporate human intelligence into the machine learning loop, primarily during the training phase and subsequent iterations, which enhances the model’s understanding and performance while gradually reducing the dependency on human input.*
    - *In the context of penetration testing, the HITL approach is crucial because it allows the AI models to learn from the nuanced decisions made by human experts. For instance, during the initial training phase, cybersecurity experts can provide insights into complex attack vectors or unusual patterns of network traffic that an AI might not initially recognize. By integrating these human judgments into the training data, the AI learns to identify similar threats autonomously, thus reducing the future need for human intervention.*
    - *Moreover, the paper highlights the role of HITL in improving data processing and annotation, which are foundational to building robust AI models. Human experts can annotate and preprocess data to create high-quality datasets that reflect the real-world complexities of cybersecurity threats. These enhanced datasets enable AI models to better learn the characteristics of sophisticated cyber-attacks and develop more effective strategies for their detection and mitigation.*
    - *As AI models trained with HITL become more adept at simulating human-like decision-making processes, their reliance on ongoing human input diminishes. This evolving independence is crucial in penetration testing, where the ability of AI models to adapt and respond to new threats quickly can significantly enhance an organization's defensive capabilities. For example, an AI model that has been trained to recognize the tactics, techniques, and procedures of various threat actors can autonomously conduct penetration tests, identify vulnerabilities, and suggest or implement countermeasures, speeding up the response time and reducing the workload on human security teams.*
    - *In conclusion, HITL systems not only foster the initial development of sophisticated AI models for penetration testing but also ensure these models remain effective as the landscape of cybersecurity threats evolves. The strategic integration of human insights allows these systems to achieve a balance where the AI's growing autonomy is continuously informed by expert knowledge, leading to reduced reliance on human input while maintaining high standards of security and adaptability.*

### Enhancing Data Accuracy

* + Data Verification Protocols:
    - Establish protocols to verify the authenticity and accuracy of input data. This includes the implementation of checksums, anomaly detection systems, and cross-validation with trusted databases.
  + Real-Time Data Correction:
    - Develop algorithms capable of real-time data correction and noise reduction to handle incomplete or biased inputs. Techniques such as Kalman filters or Bayesian networks could be applied to predict and correct errors in the data stream dynamically.
  + Synthetic Data Generation:
    - Utilize synthetic data generation techniques to enhance the training of AI models. Generative adversarial networks (GANs) can be employed to create realistic network traffic scenarios for testing and training, ensuring that the AI models can adapt to evolving network behaviors without direct exposure to sensitive or operational data.

### Increasing Model Robustness

* + Ensemble Learning:
    - Deploy ensemble learning methods to integrate multiple AI models. This approach reduces the risk associated with individual model failures and improves the generalizability across different network environments. Techniques such as random forests, gradient boosting machines, and stacking models can be utilized.
  + Regularization Techniques:
    - Implement regularization methods to prevent overfitting and enhance the generalizability of the models. Techniques such as dropout, L1/L2 regularization, and early stopping should be considered during the training phase.
  + Adversarial Training:
    - Incorporate adversarial training in the learning process. This involves training the model on perturbed datasets, which include deliberate, synthetic adversarial examples. This method helps improve the model's resilience against attacks that try to exploit model weaknesses.

### Improving Environmental Adaptability

* + Reinforcement Learning (RL):
    - Utilize RL to develop models that can dynamically adapt to new threats and environmental changes without human intervention. RL agents can be trained to optimize security decisions continuously based on rewards linked to successful threat mitigation.
  + Transfer Learning:
    - Implement transfer learning to enable models to apply knowledge gained from one domain to different but related domains. This is particularly useful in PT, where AI models trained in one network environment may need to be quickly adapted to secure another environment.
  + Hybrid AI Approaches:
    - Explore hybrid AI approaches that combine rule-based systems with machine learning models. This can facilitate better decision-making by leveraging the interpretability of rule-based systems with the adaptability of machine learning.

# Implementation and evaluation

## Simulation Design

Outline of how simulations are used to test the effectiveness of AI-driven PT tools in a controlled environment.

## Metrics for Success

Definition of the metrics that will be used to evaluate AI tools, such as detection rate, false positives/negatives, and adaptability to network changes.

## Results and Analysis

Presentation of the results from the simulation tests.

Analysis of how the AI tools performed against set metrics, discussing successes and areas for improvement.

# Comparison

a

# Conclusion and discussions

a