# Introduction

## Background of the Course Topic

* + - a

## Importance and Relevance of the Course Topic

* + - a

# Literature Review

## Research Paper 1: Introduction of Topic

### Details

* + - * Getting pwn’d by AI: Penetration Testing with Large Language Models
      * Andreas Happe, Jürgen Cito
      * [link](https://arxiv.org/pdf/2308.00121.pdf)

### Introduction

* + - * Challenges/ complexities of Penetration Testing
        + high levels of expertise
      * Necessity of for Sparring Partners
        + partners offer alternative ideas/approaches
        + “A good sparring partner should be able to cover the different tactics, techniques, and procedures covered by ATT&CK.”
        + recent interview emphasized need for human sparring partner
        + critical shortage of skilled security professionals

ISC2 Cybersecurity Workforce Study 2022

* + - * Use of Large Language Models (LLMs), in penetration testing
        + LLM is a neural network that is highly trained with an extensive volume of data using self-supervised learning techniques.
        + Emergence of Large Language Models (LLMs)

*The first thought many people have when they hear AI in relation to society today, they think of LLMs*

ie ChatGPT, GPT3.5, AutoGPT

*While the traditional GPT prompts, ChatGPT, GPT3.5, are started by a user initiating the conversation with a prompt… AutoGPT automates the initial query/prompts*

* + - * + Using AI to train “novice penetration testers”
        + Outsource sparring partnerships to AI [SEGUE]

### Summary

* + - * explores Outsource sparring partnerships to AI
        + “empower existing human security testers”
        + “counteract the lack of sufficiently educated security professionals”

#### High-Level Guidance (Task Planning)

* + - * + planning phase

tasks that involve strategic planning and decision-making

* + - * + Examples:

design overall pentest plan,

determine tactics/techniques, and

identify potential vulnerabilities

#### Low-Level Guidance (Vulnerability Hunting)

* + - * + Execution phase

More detailed/ specific actions

provide step-by-step assistance during execution

often after the strategic planning has completed

Assumed Pentesters have already selected a specific tactic and just require specific TTPs

* + - * + Examples:

Identifying/targeting vulnerabilities within a specific system

executing commands/ specific exploits

information on how to escalate privileges or perform specific attacks

### Methodologies

* + - * referred to MITRE ATT&CK to identify various tactics, techniques, and procedures (TTPs) that are commonly used by adversaries
        + Authors “assume that a sparring partner for penetration testing should cover the whole TTP spectrum”

#### High-Level Guidance (Task Planning)

* + - * + Expectation

LLM “should be able to select suitable tactics and corresponding techniques.”

* + - * + Experiment

used to provide guidance/ insights at a strategic level.

Ie suggest a penetration testing plan for a target organization/ recommend tactics to use

prompted the “LLM to help design penetration tests for both generic scenarios as well as for a concrete target organization”

generic

*To achieve this, the authors instructed AgentGPT, an autonomous AI utilizing the GPT3.5 architecture, with the task of "Becoming a domain admin in an Active Directory."*

Concrete

AutoGPT is tasked with creating a penetration testing plan to test an approved organizations external-facing systems, networks, and digital assets from the perspective of a potential external attacker.

AutoGPT also continued on to crawl the company's website and identified potential targets for phishing attacks, (users and their email addresses)

* + - * + Results

AutoGPT returns a comprehensive plan which included common/standard methods :

Network vulnerability scanning.

OSINT (Open-Source Intelligence)

user enumeration

Phishing attacks against identified users.

“declined to perform any “real” network security scan or perform phishing operations due to its ethical filters.”

“Both answers were realistic, and feasible, and would give a penetration tester good feedback about potential attack vectors.”

#### Low-Level Guidance (Vulnerability Hunting)

* + - * + Expectation

LLM “given an employed tactic, it should be able to derive feasible techniques and procedures”

The main goal is to achieve privilege escalation and gain root access on the virtual machine.

* + - * + Experiment

analyze the state of a vulnerable virtual machine, suggest specific commands or actions to exploit vulnerabilities

Scenario: Assumed Pentesters have already gained some level of access, and require guidance to escalate to root

Authors create A deliberately vulnerable Linux virtual machine

set up a connection between GPT3.5 and a VM (VM was intentionally designed to have security vulnerabilities)

“Python script that uses SSH to connect to a deliberately vulnerable *lin.security* Linux virtual machine”

Asked the LLM to analyze the VM's state, generate commands or actions, and potentially control or influence the VM's behavior.

The script operates in an infinite loop.

Within the loop, it instructs GPT3.5 (or LLM) to “imagine being a low-privilege user” and to suggest Linux shell commands.

The suggested commands are executed over SSH on the vulnerable virtual machine.

The command and resulting output from the vm are then sent to GPT and it is asked to analyze the provided command and its output and identify any potential security vulnerabilities in the system

IF GPT identifies a vuln, it is instructed to provide steps on how to exploit the vulnerability

To bypass the ethical filters, the authors utilize prompt engineering by asking for “verification commands” instead of “exploitation examples”

Note: all command, output, and subsequent actions were meticulously documented.

Script is ran multiple times

* + - * + Results

successfully obtained root privileges

Identified and exploited security vulnerabilities, including the use of sudo commands, GTFObins, and retrieval of /etc/passwd for privilege escalation.

“GTFObins are benign system commands that when called through sudo, can be abused to gain a root shell.”

Recorded and analyzed command executions and their outputs to determine vulnerability grounding.

### Main Findings

#### General

* + - * + GPT3.5 displayed signs of understanding causal relationships between actions and outcomes during the experiments.

*the suggestions provided by the LLM seemed to exhibit a degree of logical thinking or understanding of cause-and-effect relationships.*

* + - * + GPT3.5 suggestions consistently followed logically from the obtained data

LLM Was able to make suggestions based on its broader understanding of common vulnerabilities in Linux environments, even without specific information about the target system's configuration or weaknesses.

* + - * + However, the suggested system commands were clearly based on pattern-matching

*While there were instances where the LLM's suggestions might have seemed rational, the authors emphasize that these suggestions were still primarily driven by pattern recognition and did not reflect a deep understanding of the subject matter. They are acknowledging that the LLM's responses might appear reasonable on the surface but are ultimately based on data patterns rather than true comprehension.*

*despite the appearance of reasoning,… suggestions were primarily generated based on patterns it had learned from its training data, rather than a genuine comprehension of the Linux systems.*

* + - * + *The outcomes provided included "realistic, feasible, and commonly employed" exploit options*

#### Hallucinations

* + - * + Invention of “facts that seem statistically plausible.”
        + Can be reduced by “using external knowledge and automated feedback” such as with AutoGPT
        + Happened rarely and were easily identifiable

*These hallucinations were infrequent and easily identifiable, with the most common example being the suggestion to execute "exploit.sh."*

#### Reproducibility

* + - * + On a small scale, the performance of LLMs are unstable and inconsistent

variations in the commands generated and vulnerabilities identified during these runs

Variation in single runs related to focus on specific aspects

likened to the experience of humans who become overly focused on particular a element and end up “going down a rabbit hole”

* + - * + longer runs, or aggregated results from multiple runs led to more consistent outcomes
        + outcomes/ results produced were less predictable/ consistent compared to traditional enumeration tools like linpeas.sh.

“Compared to tools such as linpeas.sh [30], LLMs seem to be less deterministic”

#### Ethical Moderation in LLMs

* + - * + LLMs are limited by their ethics filters

safety measures against malicious prompts

* + - * + Prompt Engineering

Slight variations in prompts helped reduce ethical concerns.

ie asking for “verification commands” instead of “exploitation examples”,

ie instructing the AI not to ask questions or make judgments

these prompts helped reduce ethical denials and prove that these ethical filters can be bypassed

can be used by both legitimate security professionals and malicious actors.

* + - * + switching from OpenAI to one of the
        + locally running LLMs would remove all server-side ethics checks.

#### Future:

* + - * + propose enhancing the collaboration between penetration testers and AI by integrating high- and low-level tasks into a single AI system.
        + suggest evaluating locally run AI models for cost savings and customization.
        + Improving AI memory and context retention could reduce errors
        + automated prompt generation could enhance query effectiveness.
        + emphasize the need for cybersecurity professionals to prepare for AI-driven attacks

### Relevance to Your Course Content

#### Provide as an introduction to AI in pentesting

* + - * + Focuses on one of the more popular AI models people are familiar with: the integration of AI, specifically LLMs, into penetration testing.
        + Shows some basic prompt-response techniques that can be leveraged to assist in high-level planning and low-level vulnerability identification
        + This integration also shows how AI can streamline various aspects of penetration testing

#### Real-world Application:

* + - * + The paper shows practical applications of AI in penetration testing scenarios.
        + Real-world scenarios are important for allowing students to see AI's transformative power through real examples. (interesting, engaging, etc)

#### Introduction to Ethical Dilemmas:

* + - * + Begins introducing the ethical challenges/ dilemmas that arise when AI is used in penetration testing

Remember that many hackers are not concerned with the ethics

When is prompt engineering okay/when is it not?

* + - * + Explain that while AI can enhance security, it also brings about unique ethical concerns.

## Research Paper 2: Pentesting Step 1 - Gathering Information (Module 1)

### Details

* + - * Reinforcement Learning for Intelligent Penetration Testing
      * Mohamed C. Ghanem, Thomas M. Chen
      * [link](https://ieeexplore.ieee.org/document/8611595)

### Introduction

* + - * a

### Summary

* + - * A

### Methodologies

* + - * a

### Main Findings

* + - * a

### Relevance to Your Course Content

* + - * a

## Research Paper 3: Pentesting Step 2 - Scanning (Module 2)

### Details

* + - * Penetration Testing Procedure using Machine Learning
      * Reevan Seelen Jagamogan, Saiful Adli Ismail, Noor Hafizah Hassan, Hafiza Abas
      * [link](https://ieeexplore.ieee.org/document/9870951)

### Introduction

* + - * a

### Summary

* + - * a

### Main Findings

* + - * a

### Methodologies

* + - * a

### Relevance to Your Course Content

* + - * a

## Research Paper 4: Pentesting Step 3 - Exploiting (Module 3)

### Details

* + - * An Intelligent Penetration Test Simulation Environment Construction Method Incorporating Social Engineering Factors
      * Yang Li, Yongjie Wang, Xinli Xiong, Jingye Zhang, Qian Yao
      * [link](https://www.mdpi.com/2076-3417/12/12/6186)

### Introduction

* + - * a

### Summary

* + - * a

### Main Findings

* + - * a

### Methodologies

* + - * a

### Relevance to Your Course Content

* + - * a

## Research Paper 5: Post-Breach/Exploit, Pentesting Steps 4,5, and 6 (Module 4)

### Details

* + - * Automated Post-Breach Penetration Testing through Reinforcement Learning
      * Sujita Chaudhary, Austin O’Brien, Shengjie Xu
      * [link](https://ieeexplore.ieee.org/document/9162301)

### Introduction

* + - * a

### Summary

* + - * a

### Main Findings

* + - * a

### Methodologies

* + - * a

### Relevance to Your Course Content

* + - * a

# Incorporation of Findings into the Course

## Discussion on how the findings from the literature review inform your course development.

* + - a

## Preliminary Ideas for Course Content based on the Literature Review

* + - a

# Table of Contents for the Course

## Module 1: Introduction to AI and ML in Penetration Testing

### Module Overview and Brief Description

* + - * Module 1, comprised of "Introduction to AI and ML in Penetration Testing" and "Gathering Information and Reconnaissance with AI," serves as the foundation for the course. It offers participants an in-depth introduction to the role of Artificial Intelligence (AI) and Machine Learning (ML) in the field of penetration testing. This module introduces the first step of the Penetration Testing process, Gathering Information. Through these submodules, participants will gain a comprehensive understanding of key concepts, ethical considerations, and the terminology essential for the course.

### Submodule 1.1: Introduction to AI and ML in Penetration Testing

* + - * Overview of the Course
      * Overview of Penetration Testing
        + Pentest steps/phases
      * Overview of AI
      * Overview of Machine Learning
        + Deep Learning
      * Examine Ethical Considerations
        + Discuss Ethical challenges
        + Responsible/ lawful use of AI for security assessments
      * Key concepts and terminology
        + threats, vulnerability, exploits, AI, penetration testing, Machine Learning …

### Submodule 1.2: Gathering Information and Reconnaissance with AI

* + - * Introduction to Phase 1: Gathering Information
      * Use of AI-Driven Tools for Information Gathering
        + Shodan and Censys (“Asset Discovery”)
      * Machine Learning in Data Collection and Analysis
      * Ethical Considerations in Information Gathering
      * Demonstration: Gathering Information with AI

## Module 2: Machine Learning for Vulnerability Assessment

### Module Overview and Brief Description

* + - * Module 2, consisting of "Scanning and Vulnerability Assessment with ML" and "Exploiting and AI-Enhanced Techniques," focuses on utilizing Machine Learning (ML) in the context of vulnerability assessment during penetration testing. This module offers participants an in-depth exploration of Penetration Testing Phases 2, Scanning and Phase, and 3, Exploitation. By highlighting the role of AI and ML in these critical phases, participants will learn about AI-powered vulnerability scanners, data collection and preprocessing techniques, exploit development with ML, and practical vulnerability assessment exercises.

### Submodule 2.1: Scanning and Vulnerability Assessment with ML

* + - * Introduction to Phase 2: Scanning
      * Using Machine Learning for Vulnerability Assessment
      * AI-Powered Vulnerability Scanners
      * Data Collection and Preprocessing for ML
        + NLP
      * Practical Exercise: Vulnerability Assessment with ML

### Submodule 2.2: Exploiting and AI-Enhanced Techniques

* + - * Introduction to Phase 3: Exploitation
      * AI-Driven Exploitation Tools and Frameworks
      * Machine Learning for Exploit Development
      * Real-World Examples of AI-Enhanced Exploits
      * Demonstration: AI-Powered Exploitation

## Module 3: Post-Exploitation AI and ML Techniques

### Module Overview and Brief Description

* + - * Module 3, "Post-Exploitation AI and ML Techniques," delves into the advanced phases of penetration testing, Maintaining Connection, Covering Tracks, and Reporting. By highlighting the role of AI and ML in these critical post-exploitation phases, participants will gain insights into enhancing evasion techniques, streamlining reporting and documentation through AI-generated reports, and applying advanced AI-enhanced post-exploitation and privilege escalation strategies. This module equips participants with advanced skills and insights into the transformative power of AI and ML in post-exploitation scenarios, enhancing their ability to navigate and assess cybersecurity landscapes effectively.

### Submodule 3.1: Maintaining Connection, Covering Tracks, and Reporting

* + - * Overview of Maintaining Connection and Covering Tracks
      * Role of AI and ML in Evasion and Stealth Techniques
      * Reporting and Documentation with AI-Generated Reports
      * Practical Exercise: Maintaining Connection and Covering Tracks with AI

### Submodule 3.2: AI-Enhanced Post-Exploitation and Privilege Escalation

* + - * Techniques for Post-Exploitation with AI
      * AI-Driven Privilege Escalation Strategies
      * Realistic Scenario Simulations with AI
      * Demonstration:

## Module 4: Deep Learning and Advanced Techniques

### Module Overview and Brief Description

* + - * Module 4, "Deep Learning and Advanced Techniques," concludes the course, focusing on cutting-edge topics in penetration testing. Participants will explore the potential of Deep Learning, AI, and ML in advanced penetration testing techniques. This module wraps up the course by revisiting key concepts, exploring future trends in AI and ML within penetration testing, and offering additional insights and suggested references for further exploration in this dynamic field.

### Submodule 4.1: Deep Learning and Advanced Techniques

* + - * Introduction to Deep Learning
      * Deep Learning Applications in Penetration Testing
      * Advanced AI-Enhanced Techniques
      * Machine Learning for Zero-Day Exploits
      * AI-Driven Red Team Operations
      * Demonstration:

### Submodule 4.2: Review/Conclusion

* + - * Review Key Concepts
      * Future Trends in AI and ML in Penetration Testing
      * Additional Insights
        + Discuss References and Further Reading

## References and Further Reading

* + - a

# Key Components of the Newly Developed Course

## Learning Objectives

* + - *outline what students should be able to achieve or demonstrate by the end of the course*
    - Understand fundamental concepts of pentesting, AI, and ML
    - Understand penetration testing phases
    - Familiarity with various pentesting tools

## Target Audience

* + - *group of individuals or professionals for whom the course is designed*

## Assessment Strategies

* + - *how students' knowledge and skills will be evaluated and measured throughout the course*
    - A quiz at the end of each module to test for understanding
    - Assorted practical exercises

## Supplementary Materials

* + - *“description of what you intend to include in your future final deliverable”*
    - reading materials
      * e.g., online technical blogs, articles, whitepapers
      * about any topics, tools, and related concepts mentioned
    - official tutorials of some
      * software and/or hardware tools mentioned/ used

# Conclusion

## Summary of Key Insights

* + - a

## Potential Impact of the Course

* + - a

# References

# Additional Facts

* + “cybercriminals are increasingly doubling the effectiveness of their attack tools for half the cost every few months”
    - <https://ieeexplore.ieee.org/abstract/document/8963730>

# Additional Resources to Consider

* + Reinforcing Penetration Testing Using AI
    - <https://ieeexplore.ieee.org/abstract/document/9843459>
  + Reinforcement Learning for Intelligent Penetration Testing
    - <https://ieeexplore.ieee.org/document/8611595>
  + Survey On The Applications Of Artificial Intelligence In Cyber Security \*\*
    - https://www.researchgate.net/profile/Aaron-Achi/publication/355119649\_Survey\_On\_The\_Applications\_Of\_Artificial\_Intelligence\_In\_Cyber\_Security/links/623c6b5991e0810f44d62f22/Survey-On-The-Applications-Of-Artificial-Intelligence-In-Cyber-Security.pdf