How AI Will Impact Penetration Testing in Cybersecurity Measures

Daniel Carbajal

CYBR 650 Trends in Cybersecurity

M5: Capstone Project – Final Paper

August 6, 2023

Table of Contents

[I. Penetration Testing Introduction 4](#_Toc142142015)

[II. Penetration Testing Standards 5](#_Toc142142016)

[A. Open-Source Security Testing Methodology Manual 5](#_Toc142142017)

[B. NIST Special Publication 800-42 5](#_Toc142142018)

[C. Penetration Testing Execution Standard 6](#_Toc142142019)

[D. Open Web Application Security Project 6](#_Toc142142020)

[III. Penetration Testing Process 6](#_Toc142142021)

[A. Pre-engagement Interactions 6](#_Toc142142022)

[B. Information Gathering 7](#_Toc142142023)

[C. Threat Modelling 7](#_Toc142142024)

[D. Vulnerability Analysis 7](#_Toc142142025)

[E. Exploitation 8](#_Toc142142026)

[F. Post Exploitation 8](#_Toc142142027)

[G. Reporting 8](#_Toc142142028)

[IV. Penetration Testing Tools 9](#_Toc142142029)

[A. Penetration Testing Platform: Kali Linux 9](#_Toc142142030)

[B. Information Gathering: Nmap 9](#_Toc142142031)

[C. Vulnerability Scanner: Nessus and OpenVAS 9](#_Toc142142032)

[D. Exploitation: Metasploit, Core Impact, and CANVAS 10](#_Toc142142033)

[E. Password Attack Tools: Hydra and John the Ripper 11](#_Toc142142034)

[F. Web Security Assessment Framework: W3af and Sqlmap 11](#_Toc142142035)

[G. Man-in-the-Middle Attack Tool: Ettercap 11](#_Toc142142036)

[H. Social Engineering Attack Tool: SET 12](#_Toc142142037)

[V. What does Penetration testing involve? 12](#_Toc142142038)

[A. Black box penetration testing 12](#_Toc142142039)

[B. White box penetration testing 13](#_Toc142142040)

[C. Gray box Penetration testing 13](#_Toc142142041)

[VI. The Penetration testing teams 14](#_Toc142142042)

[A. The Red Team 14](#_Toc142142043)

[B. The Blue Team 14](#_Toc142142044)

[C. The Purple teams. 14](#_Toc142142045)

[VII. The types of Penetration tests 15](#_Toc142142046)

[VIII. Penetration testing Using Artificial Intelligence 15](#_Toc142142047)

[IX. What Data Should be Used for Machine Learning Portion of Penetration Testing 17](#_Toc142142048)

[X. Classification of Attacks 18](#_Toc142142049)

[A. Information Gathering 18](#_Toc142142050)

[B. Configuration Error Attack 19](#_Toc142142051)

[C. Buffer Overflow Attack 19](#_Toc142142052)

[D. Password Attack 20](#_Toc142142053)

[E. Web Attack 20](#_Toc142142054)

[F. Sniffer Attack 21](#_Toc142142055)

[G. Social Engineering Attack 21](#_Toc142142056)

[H. Denial of Service Attack 22](#_Toc142142057)

[XI. How AI helps cybersecurity governance 22](#_Toc142142058)

[XII. How AI helps PT and risk mitigation 23](#_Toc142142059)

[XIII. How AI helps PT and continuous monitoring and reporting 24](#_Toc142142060)

[XIV. Conclusion 24](#_Toc142142061)

[XV. References 25](#_Toc142142062)

# Penetration Testing Introduction

Penetration testing (PT) is an established approach to evaluating the security of digital assets by actively identifying and exploiting existing vulnerabilities. It is an approach used to improve the information security level of the target system. However, the use of PT has been restricted to advanced security experts with many years of experience (He & Bode, 2006**)**. Furthermore, the complex manual process is costly and time-consuming.

Automation can significantly reduce the time, cost, and human labor required for information gathering, analysis, and exploitation. In terms of privacy protection, automated PT can prevent human testers' leakage of sensitive information. Adding automation can also extend the field of Artificial Intelligence (AI). The automation you can achieve through artificial intelligence could help make PT much easier to do consistently and at scale. Machine learning (ML) is one of the hottest areas in data science (Frąckiewicz,2023). This subset of artificial intelligence allows a system to learn from data and make accurate predictions, identify anomalies, or make recommendations using different techniques. Machine learning techniques extract information from vast amounts of data and transform it into valuable business knowledge. While most industries use these techniques, they are especially prominent in the finance, marketing, healthcare, retail, and cybersecurity sectors.

Machine learning can also address new cyber threats. Many types of cyberattacks include structured query language (SQL) injection, phishing, cross-site script attacks, malware, social engineering, man-in-the-middle attacks, distributed denial of service attacks, and ransomware (Frąckiewicz,2023). Organizations employ machine learning to constantly evaluate data, find patterns that could result in potential attacks, and mitigate them.

The rest of this paper thoroughly covers PT and how machine learning can be leveraged to automate and simplify cybersecurity. It can help organizations create a culture that can tackle the skills gap and provide security personnel with mature testing strategies, allowing them to protect better the organization’s information technology (IT) assets and information.

# Penetration Testing Standards

PTs are highly complex tasks and techniques requiring participants to have a relatively high level of skills as they engage in various complex scenarios. Following are standards for the PT methods, processes, and steps in the information security field.

## Open-Source Security Testing Methodology Manual

The Open-Source Security Testing Methodology Manual (OSSTMM) was published by the Institute for Security and Open Methodologies (ISECOM) (Herzog, 2003). It is a popular international standard for information security testing and analysis in many organizations. It covers all the elements of PT, including physical security, psychology, data networks, wireless communication, and telecommunications facilities.

OSSTMM can significantly reduce false negatives and positives and provide more accurate security metrics. Among OSSTMM’s more important features are its incredible attention to technical details and good operability.

## NIST Special Publication 800-42

The US Government’s National Institute of Standards and Technology (NIST) has published the 800-series special publications (SP) defining IT security governance and controls. (Wack et al., 2003) introduces security testing techniques, system development life cycles, strategies, and standard testing tools.

## Penetration Testing Execution Standard

The Penetration Testing Execution Standard (PTES) (Nickerson et al., 2014) is a relatively new standard developed in 2010 by information security experts. It defines a practical PT process that consists of seven stages. Moreover, it is a comprehensive PT framework that covers all the technical aspects of PT, including expert experience and related tools. PTES is one of the most famous PT standards in the information security industry.

## Open Web Application Security Project

The Open Web Application Security Project (OWASP) is a non-profit organization focusing on web security, providing security testers and developers with guidelines to identify and avoid security threats. Each year, OWASP publishes a top 10 (Andrew et al., 2021) threats security report covering the most common security issues on web applications. These reports are widely used and analyzed in detail by information security experts.

# Penetration Testing Process

PT Testing has the following seven stages and has been widely accepted by the security industry (Nickerson et al., 2014). Each stage is discussed in further detail in the following sub-sections.

## Pre-engagement Interactions

In the pre-engagement interactions stage, the PT team discusses test technology, test target, test scope, test cycle, test scheme, and the corresponding price with clients. In general, PT should not affect the availability of the target.

## Information Gathering

After the pre-engagement interactions stage, the PT team must learn about the targets. Information gathering is one of the most critical stages in PT. It aims to collect as much information as possible, such as physical information, logical relationships, organizational structure, physical assets, individual information, footprinting information, and protection mechanisms (Nickerson et al., 2014). The more information collected during this stage, the more attack vectors may be used.

## Threat Modelling

After the information-gathering stage, the PT team conducts threat modeling and attack planning to determine the most feasible attack path based on the information obtained. Threat modeling consists of business asset analysis, business process analysis, threat agents/community analysis, threat capability analysis, motivation modeling, and finding relevant news of comparable organizations being compromised (Nickerson et al., 2014). Regarding attack planning, PT teams determine the attack methods, tools, and schemes.

## Vulnerability Analysis

Vulnerability analysis is a process of discovering vulnerabilities in systems and applications (Nickerson et al., 2014). The PT team needs to appropriately consider the scope of testing for the depth and breadth of applications to meet the desired outcome's goals and requirements. The vulnerability analysis process includes active testing, passive testing, validation, and research. Sometimes, experienced teams can find zero-day (unknown) vulnerabilities in target systems.

## Exploitation

The exploitation stage is the most challenging part of PT. The PT team performs attacks on targets, such as SQL injection, password, buffer overflow, cross-site script (XSS), man-in-the-middle (MITM), and social engineering attacks (Nickerson et al., 2014). Typically, the targets are protected by different kinds of countermeasures such as anti-virus, intrusion detection system (IDS), web application firewall (WAF), packing, cryptography, white-black list, data execution prevention (DEP), and address space layout randomization (ASLR). Thus, the exploitation stage focuses on performing a successful attack by bypassing security countermeasures in the target system (Nickerson et al., 2014). Moreover, in the case of Black-box testing, the PT team must avoid being discovered by the target security team.

## Post Exploitation

The purpose of the post-exploitation stage is to keep control of the machine for future use. In this stage, the PT team analyses network interfaces, routing, domain name system (DNS), cache ARP tables, proxy servers, network services, and directory information to identify other targets for further attack and install backdoor programs to maintain the long-term access privilege of a target (Nickerson et al., 2014). A clean-up process is sometimes applied to systems once the PT has been completed.

## Reporting

Finally, after the execution of the first six stages, a report is submitted to the client for the entire task, which outlines all aspects of PT, such as objectives, methods, and results, and gives repair solutions. The report generally includes a PT and technical summary (Nickerson et al., 2014).

# Penetration Testing Tools

Various tools or frameworks are available in each PT stage to perform information gathering and different kinds of attacks. This section introduces some of the essential penetration tools.

## Penetration Testing Platform: Kali Linux

Kali Linux (Allen et al., 2014) is a Debian-based Linux distribution aimed at advanced PT and security auditing, which Offensive Security maintains and funds. Kali Linux contains over 600 PT tools for various information security tasks, such as PT, security research, computer forensics, and reverse engineering. Kali Linux is specifically designed to meet the needs of PT professionals.

## Information Gathering: Nmap

Nmap (Lyon, 2009) is the best-known and most professional security scanner and can be used to discover network ports, hosts, and services. It was written in C/C++ and Python by Gordon Lyon starting in 1997. To discover hosts on a network, Nmap sends specially built packets to the target host and then analyses responses. The program is different from other available port scanners. Nmap sends packets based on network conditions. Unlike other scanners, Nmap can scan ports, discover online hosts, and recognize the system type running in remote hosts. In general, Nmap is an essential tool in the information-gathering stage.

## Vulnerability Scanner: Nessus and OpenVAS

A vulnerability scanner is a program that automatically finds and discovers security vulnerabilities in computers, information systems, networks, and applications. It identifies vulnerabilities by sending specific packets to the target and then analyzing responses to match its vulnerability database. Nessus (Beale et al., 2004) is the world’s most famous vulnerability scanner, used by over 75,000 organizations worldwide. The tool provides a full vulnerability scanning function and frequently updates its vulnerability library. Like Nessus, OpenVAS (Aksu et al., 2019) is an open-source branch of the Nessus project and one of the most popular vulnerability scanners. In the information-gathering stage, a vulnerability scanner is the best way to discover known vulnerabilities in the target system.

## Exploitation: Metasploit, Core Impact, and CANVAS

Metasploit (Kennedy et al., 2011) is the most famous PT framework. It provides tools to be exploited against remote targets and contains hundreds of professional exploit tools for known software vulnerabilities. Before Metasploit was published, penetration testers had to repeat the complex process of exploiting a code search, compiling, testing, modifying exploit code, execute the exploit until they achieved success. Metasploit collects exploits and allows users to develop exploits in their environment.

Core Impact (Ferreira & Kleppe, 2011) is an expensive commercial PT system developed by Core Security Technologies. It enables security teams to exploit security weaknesses, increase productivity and improve efficiency. Core Impact is designed for users at every level, from beginners to experts, and all modules, exploits, and tools are written in Python. It includes professional exploit libraries and engines that perform PT on web applications, network systems, user terminals, and wireless networks.

CANVAS (He and Bode, 2006) includes hundreds of exploits and is an automated exploitation system for penetration testers and security professionals worldwide. Moreover, it is a platform designed to allow the easy development of other security products. Immunity, the company which developed CANVAS, also provides services, products, and education around information security.

## Password Attack Tools: Hydra and John the Ripper

Hydra is a powerful online password attack tool that can support most protocols or applications, such as FTP, HTTP, HTTPS, MySQL, MSSQL, Oracle, Cisco, IMAP, and VNC (Mello, 2023). John the Ripper is a famous password attack tool in the Linux system (Mello, 2023). The success rate of password cracking is related to the dictionary.

## Web Security Assessment Framework: W3af and Sqlmap

W3af is a widely used web application attack and audit framework (Riancho, 2011). The project aims to create a framework to help administrators secure their web applications by finding and exploiting all vulnerabilities. This framework is developed using Python; thus, it is easy to use and extend. W3af can identify more than 200 vulnerabilities in web applications, including SQL injection, XSS, guessable credentials, and unhandled application and PHP configuration errors.

Sqlmap (Damele & Stampar, 2012) is another web attack tool that automates detecting and exploiting SQL injection. It has a powerful engine that automates the following operations: (I) database Identification, (II) obtaining data from the database, (III) accessing the underlying file system, and (IV) executing commands on the operating system.

## Man-in-the-Middle Attack Tool: Ettercap

A MITM attack is performed through data tampering and sniffing attacks by intercepting communication data in a target network (Callegati et al., 2009). Usually, MITM attacks are challenging to detect. Ettercap (Norton, 2004) is a comprehensive suite for MITM attacks, which can be used for computer network protocol analysis and security auditing. It features, among other elements, sniffing live connections and content filtering. Ettercap supports the active and passive dissection of many protocols.

## Social Engineering Attack Tool: SET

Social engineering attack (Mouton et al., 2016) is an attack vector that relies heavily on human interaction and often involves manipulating people into breaking standard security procedures and best practices to gain access to systems, networks, or physical locations or for financial gain. In high-level PT, targets are often well protected; thus, social engineering attacks are often the key to success for the attacker. SET (Conheady, 2014) is the best-known social engineering tool and can perform 11 social engineering attacks.

# What does Penetration testing involve?

To uncover the vulnerabilities which can be found in type or kind of Web Application, three types of PT can be used (Das, 2019), which are as follows:

* Black Box Testing.
* White Box Testing.
* Gray Box Testing.

## Black box penetration testing

In a real-world cyber-attack, the hacker probably will not know all the ins and outs of the IT infrastructure of a corporation. Because of this, they will launch a brute-force attack against the IT infrastructure, hoping to find a vulnerability or weakness to exploit.

In other words, in this type of PT, there is no information given to the tester about the internal workings of the network, software applications, and security infrastructure (Das, 2019). As a result, this test can take a very long time to complete, so the tester will often rely upon automated processes to uncover weaknesses and vulnerabilities. This test type is also called the “trial and error” approach.

## White box penetration testing

In this type of PT, the tester has full knowledge and access to the network and security applications (Das, 2018). Because of this, a White Box Test can be accomplished much faster when compared to a Black Box Test. The other advantage is that a much more thorough PT can be completed because the tester can focus on more internal weaknesses and vulnerabilities rather than trying to gain access and navigate around the network. However, this approach also has its set of disadvantages. First, since a tester has complete knowledge, deciding what to focus on regarding system and component testing and analysis could take more time. Second, more sophisticated tools, such as software code analyzers and debuggers, are required to conduct this type of test.

## Gray box Penetration testing

As the name implies, this type of test combines the Black Box and the White Box Test. In other words, the penetration tester only has partial knowledge of the internal workings of the network and its applications (Das, 2019). This is often restricted to accessing the software code and system architecture diagrams.

Manual and automated testing processes can be utilized with the Gray Box Test. Because of this approach, a PT tester can focus their primary efforts focus on those areas of the network and applications that the PT tester knows most about and exploit any weaknesses or vulnerabilities (Das, 2018). With this method, there is a higher probability that more difficult “security gaps” will also be discovered.

# The Penetration testing teams

Very often, when it comes to PT, the image of just one person doing the test is conjured up. However, keep in mind that the best types of PT come into play when multiple testers are utilized and are broken down into three teams (Das, 2018), which are as follows:

* The Red Team
* The Blue Team
* The Purple Team

## The Red team

The Red Team can be considered as those individuals who are the actual PT testers. Their primary goal and objective are to mimic or emulate the mindset of an attacker, trying to break down all the weaknesses and vulnerabilities present (Das, 2018). In other words, the Red Team attacks all fronts possible.

## The Blue team

The Blue Team can be considered personnel from within the business’s infrastructure. This can be the IT Security team, whose primary goal and objective are to thwart and defend against any attacks from the Red Team. Anybody participating on the Blue Team must possess the mindset of constant proactiveness and vigilance to defend the corporation against all attacks (Das, 2019).

If you think about it, the Red and Blue Team can be viewed as the two sides of a particular coin. The summation goal of these two teams is to constantly enhance the corporation's security posture by sharing feedback. However, this does not always happen. Thus, there is a need for the Purple Team.

## The Purple teams.

The Purple Team can be viewed as the composite of the Red and Blue Teams. The Purple Team adopts the security controls and tactics from the Blue Team, as well as the security weaknesses and vulnerabilities which the Red Team discovers (Das, 2018). This is then all translated into a single narrative that can be shared across all the teams to implement a policy of continuous and constant security improvements for the organization.

# The types of Penetration tests

Now that the teams have been divided and their roles and responsibilities clearly defined, some types of PT can be engaged. (Broad & Bindner, 2013) These are as follows:

* Network Services
* Web Application
* Client Side
* Wireless
* Social Engineering

# Penetration testing Using Artificial Intelligence

A natural question arises regarding the capability of AI to provide a potential solution beyond simple automation to achieve expert-like output (Creasy & Glover, 2019). In other research fields, AI has proven helpful in offloading work from humans and possibly handling depths and details that humans cannot tackle quickly or accurately (Spaan, 2012). Rapid progress in AI and, notably, the machine learning (ML) sub-field led us to believe that an AI-based PT system utilizing well-grounded models and algorithms for making sequential decisions in uncertain environments can bridge the gap between automation and expertise that the PT community experience (Hoffman, 2015). In this perspective, the existing PT systems and framework started shifting from executing experts’ tasks to becoming more autonomous, intelligent, and optimized, aiming that all existing threats are checked systematically and efficiently without or with little human expert intervention (Sarraute, 2019). Furthermore, these systems should optimize the use of resources by eliminating time-consuming and irrelevant directions and ensuring no threat is overlooked.

In addition to the regular use of PT, the testing results (output) should be processed and stored to serve for further use (Spaan, 2012). The main difference between human PT experts and automated systems is that humans learn alongside performing the tests and enrich their expertise throughout, while systems omit the re-usability of the data, which is sometimes crucial, especially when the testing is repeated, such as regular compliance tests (Creasy & Glover, 2019). In practical terms, most of the assessed network configurations will not change considerably over a short period.

Therefore, the output of previous tests could remain entirely or partly applicable for an eventual re-testing required after one or more of these following points occur:

• Network hardware, software, segments, or applications were added, changed, or removed

• Significant systems upgrades or modifications are applied to infrastructure or applications

• Infrastructure changes, including moving locations

• Security solutions or patches were installed or modified

• Security or users policies were modified

Automation is the best solution to save time and resources in any domain, and PT is not an exception to this rule. Therefore, the offensive cybersecurity community has given particular attention to automation during the last decade, leading to improvements in saving significant time, effort, and resources in performing tasks (Almubairik & Wills, 2016). Given the particularity of PT practice, the increasing size and complexity of the tested assets, and the significant number of vulnerabilities, exploits. Attack vectors that the tester should cover, the blind automated system becomes powerless. It often performs worse than manual practice pushing the researcher to focus on improving such systems by adopting various solutions.

# What Data Should be Used for Machine Learning Portion of Penetration Testing

Early research focused on improving the PT system by optimizing the planning phase, which was modeled as attack graphs or decision trees problem, reflecting the nature of PT practice as sequential decision making. Most of the works were nonetheless relevant to vulnerabilities assessment (VA) rather than PT because of the static nature of the proposed approach and its limitation to the planning phase (Qiu et al., 2014). Amongst the most significant contributions, we find the modeling of VA as attack graphs in atomic components (actions), pre-condition, and post-condition to narrow the targeted vulnerability. However, this approach was more of applying classical planning methods to find the best attack graph. Further similar works were carried out on automating the planning of PT tasks. However, blind automation did not address the problem of enhancing performance and only covered the planning phase of PT practice (Sarraute, 2019).

Nevertheless, a remarkable work on optimization was introduced by (Obes et al., 2013) by modeling PT as planning domain definition language (PDDL), which for the first time accounted for attacking and post-attacking phases of PT in addition to the flexibility offered by the solution which enabled integration with some PT systems.

Some research also considered AI to improve PT practice (Sarraute, 2019). However, most of the proposed modeling approaches failed to deal with the persistent uncertainty in PT practice, especially the lack of accurate and complete knowledge about the assessed systems. An exception was the use of ML algorithms within a professional PT and VA system called Core-Impact, in which the PT planning phase was modeled as a partially observable Markov decision process (POMDP) solved using an external POMDP solver to determine the best testing plan in form for attack vectors. However, the proposed model itself is questionable as it obviously fails to model the full PT practice and thus cannot cover the remaining testing phases and tasks, especially the vulnerability assessment, testing, and pivoting phases known to be highly interactive, sequential, and non-standard compared with the planning and information-gathering phases (Ghanem & Chen, 2018).

# Classification of Attacks

According to PT in practice (Broad & Binder, 2013), there are many types of attacks, such as information-gathering attacks, configuration attacks, buffer overflow attacks, password attacks, web attacks, sniffer attacks, social engineering attacks, and denial-of-service (DOS) attack. Along with this, an overview of how AI relates to generating these attacks is summarized as well.

## Information Gathering

Information gathering is the most critical step in PT. Typically, the target information to be collected includes IP address, open ports, application, OS type, human or organization information, network topology, defense mechanism, configuration, vulnerability, and physical environment (Broad & Binder, 2013). The collection of the above information determines whether the PT will be successful or not.

AI and ML can help the PT tester gather all the information automatically, analyze it, and determine different courses of action. For example, it can determine the best social engineering attack to deploy based on the information collected (social engineering is the use of deception to manipulate people into disclosing confidential or personal information that can be used for fraudulent purposes) (Tsukerman, 2019). Alternatively, it could be used to identify the target hosts that should be attacked first since there is a higher probability of success.

## Configuration Error Attack

This type of attack is usually based on an administrator’s system configuration error. For example, the robot.txt file usually exposes the structure information of the website, or the directory that allows users to upload files has executable permission so attackers can upload and execute a malicious file. (Herzog, 2003).

AI and ML can help by configuring several attack files to overcome the administrator’s configuration error. This can try to overwrite the root directory or the website to see if a configuration error attack is possible (Broad & Binder, 2013),

## Buffer Overflow Attack

A buffer overflow is a typical software coding mistake an attacker could exploit to access the target system (Kupperman et al., 2005). While writing data to a buffer, a program overruns the buffer’s boundary and overwrites adjacent memory locations. It allows attackers to change the program flow and execute their commands or programs. Buffer overflow is a widespread and dangerous vulnerability in many operating systems and application software. It is a famous attack used in PT.

Artificial intelligence has significant consequences for buffer overflow vulnerability. It is a driver of greater complexity, as AI systems often involve various components and algorithms that make detecting and mitigating overflow problems more challenging. Many apps nowadays integrate AI, which means the increased complexity is not just limited to a few classes of applications. However, AI itself can be used to detect buffer overflow problems. Data on how buffer overflow can be collected and a model can be devised to predict an anomaly or when someone is trying to perform a buffer overflow (Kupperman et al., 2005). This may even stem from a problem that the software application itself may create.

## Password Attack

Password attack is an essential part of PT. Usually, an attacker can gain specific permission from the target system if a password attack is successful. Most password attacks are based on a dictionary comprising possible passwords (Mello, 2023).

Based on publicly disclosed information, AI uses techniques like rainbow table attacks rather than brute forcing a password, observed Dustin Childs, head of threat awareness at Trend Micro’s Zero Day Initiative. Hackers use rainbow tables to translate hashed passwords into plaintext (Mello, 2023).

The rainbow table allows the AI to do simple searches and compare operations on a hashed password rather than a slower brute-force attack (Mello, 2023). Rainbow table attacks have been acknowledged for years and have been shown to crack even 14-character passwords in under five minutes. Older hashing algorithms such as MD5 and SHA-1 are also more susceptible to these attacks.

## Web Attack

A web attack is an attack against web applications. The most common attacks are injection, XSS, and cross-site request forgery (CSRF). The OWASP publishes a top 10 vulnerabilities yearly to raise awareness amongst developers and managers.

AI can be developed to work over recon and scanning and then perform exploitation techniques on the website (Segal, 2022). A vulnerability can be exploited in a thousand ways, but AI can be used to find the way which has the highest impact & severity. The severity of the vulnerabilities is represented in different terms depending upon the organization. OWASP summarizes the vulnerabilities from A0 to A10 according to their severity, and an AI program can make the attack vectors based on these vulnerabilities.

## Sniffer Attack

If a target system has no known vulnerabilities, an experienced human penetration tester typically attempts to perform a sniffer attack. They first break into other systems under the same sub-network with the original target, after which they monitor and then analyze all network flow to gain sensitive information such as a password (Segal, 2022).

AI can run a script through a sniffer tool like NMAP xml file, extracting information such as port number, state, server, software, and version (Tsukerman, 2019). The information fed to an ML algorithm is the port number, software, and version, and getting it to list the missing patches and security vulnerabilities related to it. It can try to exploit the vulnerabilities and see if it can break in.

## Social Engineering Attack

Social engineering attacks are directed against humans, such as administrators or users, who have weak security awareness. Social engineering refers to various malicious activities carried out through human interactions. In remote PT, these attacks are usually performed using spear-phishing attacks by emails or links, website forge attacks, or spoofing attacks (Segal, 2022).

AI algorithms can analyze social media posts, emails, and other online activity to create a profile of the target, making the attack more believable. Chatbots can be used to simulate conversations with humans. AI can allow PT testers to generate highly targeted links, websites, emails, and social media posts that people will likely click on. AI uses natural language processing to understand and interpret human language (Segal, 2022). PT testers can use AI to gather information on their targets, such as matching the victim’s profile photo across platforms to identify their various social media accounts (Tsukerman, 2019).

## Denial of Service Attack

In a DoS attack, the attackers attempt to prevent legitimate users from accessing a service. In this case, the attacker usually sends excessive data flow to the network or server to exhaust target resources. DoS attacking is not typically used in PT and usually leads to the reboot of the target system for some purpose. This attack includes SYN flood, TCP/UDP, SMTP, and ICMP attacks (Herzog, 2003). If the attack source comes from a different device, it is a distributed denial-of-service attack (DDoS) attack.

PT testers have adopted sophisticated artificial intelligence (AI) and machine learning methods to help conduct their DDoS attacks (Herzog, 2003). For example, DDoS botnets apply machine learning methods to conduct sophisticated network reconnaissance to find the most vulnerable systems. They also use AI to reconfigure themselves to thwart detection and change attack strategies (Tsukerman, 2019). Modern attacks will likely manifest as defenders and attackers pit AI-enabled systems against each other.

# How AI helps cybersecurity governance

AI can aid PT in cybersecurity governance in several ways, such as automating repetitive tasks that would otherwise take much time and resources for human PT, such as scanning, testing, and reporting vulnerabilities (Parisi, 2019). AI can also enhance the accuracy and efficiency of PT by using machine learning algorithms to analyze large amounts of data, identify patterns, predict behaviors, and find the best actions to mitigate risks. This can also allow for the adaptation to the evolving threat landscape using AI to learn from new attacks, update security models and respond to threats almost immediately. It complements human expertise and creativity by using AI to assist human PT in finding complex or hidden vulnerabilities, generating new attack scenarios, and validating security policies (Parisi, 2019).

However, AI also poses challenges for cybersecurity governance, such as ensuring the security and privacy of the data used to train and deploy AI models. The AI models themselves from malicious attacks or unauthorized access and establishing ethical principles and policies to guide the use of AI across the business and align it with corporate values, regulations, and laws and managing the trust and transparency of the AI systems and their outcomes by explaining how they work, what they do and why they do it (Marchetti,2022).

# How AI helps PT and risk mitigation

AI can aid PT in risk mitigation by helping organizations identify and prioritize the potential harms that could result from AI deployments, such as data breaches, bias, errors, or cyberattacks, providing tools and processes to control and govern AI systems by ethical principles, regulations, and best practices, such as data protection, transparency, robustness, safety, and security (TraceSecurity, 2019). It can enable faster and more effective responses to AI incidents by using AI to monitor, detect, and remediate any anomalies, vulnerabilities, or threats in real-time. Thus AI can help risk mitigation by supporting continuous improvement and learning from AI experiences by using AI to collect and analyze feedback, measure performance, and optimize outcomes (Marchetti, 2022).

# How AI helps PT and continuous monitoring and reporting

AI can aid PT in constant monitoring and reporting by automatically scanning for code and environment vulnerabilities whenever there are new assets or changes, continuously monitoring the security posture of the systems, and providing regular feedback on the status, trends, and risks (Marchetti, 2022). It can be used in integrating with the continuous integration and continuous deployment (CI/CD) pipeline to launch new scans and tests whenever there are code updates and ensure that security is embedded in the development process (TraceSecurity, 2019). Thus, AI can provide on-demand access to human experts to perform customized and threat-aware PT results as needed.

# Conclusion

AI and PT are two fields that can be closely interrelated and significantly impact the security of IT systems. AI can enhance PT's efficiency, accuracy, and adaptability by automating tasks, analyzing data, learning from attacks, and assisting human experts. PT can help to identify and mitigate the risks of AI used by adversaries by exposing vulnerabilities, testing resilience, and ensuring compliance.

However, AI and PT also face some challenges, such as ensuring the security and privacy of the data and models used for AI and PT. There is also the need to establish ethical principles and policies to guide the use of AI and PT. and manage the trust and transparency of the AI and PT systems and their outcomes. It will be necessary to use AI with PT to keep up with the evolving threat landscape and the rapid development of technologies. The future of AI and PT will depend on how well these challenges are addressed and how well these fields can collaborate and continue to complement each other.

# References

Aksu, M., U., Bicakci, K., and Altuncu, E., (2019). A first look at the usability of Openvas vulnerability scanner. In Workshop on Usable Security (USEC) 2019, NDSS.

Almubairik, N., & Wills, G. (2016). Automated penetration testing based on a threat model. In Proceedings of the 11th

International Conference for Internet Technologies and Secured Transactions, ICITST, Barcelona, Spain.

Allen, L., Heriyanto, T., and Ali, S., (2014). Kali Linux–Assuring security by penetration testing. Packt Publishing Ltd.

Andrew, S., Glas, B., & Gigler, T. (2021). Owasp top 10 2021. The Ten Most Critical Web Application Security Risks.

Beale, J, Meer, H., van der Walt, C., and Deraison, R., (2004). Nessus Network Auditing: Jay Beale Open-Source Security Series. Elsevier.

Broad, J., and Bindner, A., (2013). Hacking with Kali: practical penetration testing techniques. Newnes.

Callegati, F., Cerroni, W., and Ramilli, M., (2009). Man-in-the-middle attack to the https protocol. IEEE Security & Privacy, 7(1):78–81.

Chan, A., Can ai be used for risk assessment? ISACA, Industry News. [Can AI Be Used for Risk Assessments? (isaca.org)](https://www.isaca.org/resources/news-and-trends/industry-news/2023/can-ai-be-used-for-risk-assessments)

Conheady, S., (2014). Social engineering in IT security: Tools, tactics, and techniques. McGraw-Hill Education Group.

Creasey, J. & Glover, I. (18 December 2019). A Guide for Running an Effective Penetration Testing Program; CREST Publication: Slough,

UK, (2017). Available online: http://www.crest-approved.org.

Damele, B., and Stampar, M., (2012). Sqlmap: automatic sql injection and database takeover tool.

Das, R., (4 September 2019). The types of penetration testing, Infosec.

Das, R., (30 August 2018). How are penetration teams structured, Infosec?

Dsouza, M., (21 October 2018). How artificial intelligence can improve PT, packthub. [How artificial intelligence can improve PT | Packt Hub (packtpub.com)](https://hub.packtpub.com/how-artificial-intelligence-can-improve-pentesting/)

Fernandez, M., B., & Guarang, J., (31 May 2023). Machine learning applications in cybersecurity. Security Intelligence. [Machine Learning Applications in the Cybersecurity Space (securityintelligence.com)](https://securityintelligence.com/posts/machine-learning-applications-in-the-cybersecurity-space/?c=Artificial%20Intelligence)

Frąckiewicz, M., (16 June 2023). The impact of artificial intelligence on cybersecurity compliance, TS2 Space.

Ferreira, A., M., and Kleppe, H., (2011). Effectiveness of automated application penetration testing tools.

Garrad, & Unnikrishnan, S. (2023). Reinforcement learning in VANET penetration testing. *Results in Engineering*, p. *17*, 100970–. <https://doi.org/10.1016/j.rineng.2023.100970>

Ghanem, M. and Chen, T. (2018). Reinforcement Learning for Intelligent Penetration Testing. In Proceedings of the WS4, the World Conference on Smart Trends in Systems, Security and Sustainability, London, UK.

Grady, J. (2021). How ai benefits network detection and response, Tech Target Security, [How AI benefits network detection and response | TechTarget](https://www.techtarget.com/searchSecurity/opinion/How-AI-benefits-network-detection-and-response)

He, L., and Bode, N., (2006), Network penetration testing. In EC2ND 2005, pages 3–12. Springer.

Kennedy, D., O’gorman, J., Kearns, D., and Aharoni, M., (2011). Metasploit: the penetration tester’s guide. No Starch Press.

Herzog, P., (2003). Open-source security testing methodology manual, Institute for

Security and Open Methodologies (ISECOM).

Hilb, M., (21 July 2020). Toward artificial governance? The role of artificial intelligence in shaping the future of corporate governance. Springer Link. [Toward artificial governance? The role of artificial intelligence in shaping the future of corporate governance | SpringerLink](https://link.springer.com/article/10.1007/s10997-020-09519-9)

Hoffmann, J. (11 June 2015). Simulated penetration testing: From Dijkstra to aaTuring Test++. In Proceedings of the 25th International Conference on Automated Planning and Scheduling, Israel.

Kuperman, B., A., Bradley, C., E., Ozdoganoglu, H., Vijaykumar, T., and Jalote, A., (2005). Detection and prevention of stack buffer overflow attacks. Communications of the ACM, 48(11):50–56.

Lyon, G., F., (2009). Nmap network scanning: The official Nmap project guide to network discovery and security scanning. Insecure.

Marchetti. (2012). Enterprise risk management best practices from assessment to ongoing compliance (1st edition). Wiley.

Mello, J., P., (26 April 2023). Researchers Instantly Crack Simple Passwords With AI, Tech News World.

Mouton, F., Leenen, L., and Venter, H., S., (2016). Social engineering attack examples, templates, and scenarios. Computers & Security, 59:186–209.

Nickerson, C., Kennedy, D., Smith, E., Rabie, A., Friedli, S, Searle, J., Knight, B., Gates, C., McCray, J., Penetration testing execution standard, 2014.

Norton, D., (2004) An ettercap primer. SANS Institute InfoSec Reading Room, 5.

Obes, J.; Richarte, G.; Sarraute, C., (2013), Attack planning in the real world., arXiv:1306.4044.

Parisi. (2019). Hands-On Artificial Intelligence for Cybersecurity: Implement Smart AI Systems for Preventing Cyber Attacks and Detecting Threats and Network Anomalies. Packt Publishing, Limited.

Pentoma (23 July 2022), The launch of our ai-powered PTing solution, pentoma.blog, [The Official Launch Of Our AI-Powered PTing Solution - APP SECURITY INSIGHTS](https://blog.se.works/2018/02/27/the-official-launch-of-our-ai-powered-pen-testing-solution/)

Poston. (2022). Python for cybersecurity : using python for cyber offense and defense. John Wiley and Sons.

Qiu, X.; Jia, Q.; Wang, S.; Xia, C.; Shuang, L., (2014) Automatic generation algorithm of penetration graph in penetration testing. In Proceedings of the Ninth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, Guangdong, China.

Riancho, A., (2011), w3af-web application attack and audit framework. World Wide Web electronic publication, 21.

Sarraute, C., (20 December 2019), Automated attack planning. Available online: <https://arxiv.org/abs/1307.7808>.

Security Technologies and Methods for Advanced Cyber Threat Intelligence, Detection and Mitigation. (2022). Now Publishers. <https://doi.org/10.1561/9781680838350>

Segal, E., (2022). The impact of ai on cybersecurity, IEEE Computer Society, IEEE. [The Impact of AI on Cybersecurity | IEEE Computer Society](https://www.computer.org/publications/tech-news/trends/the-impact-of-ai-on-cybersecurity)

Spaan, M. (2012). Partially Observable Markov Decision Processes, Reinforcement Learning: State of the Art; Springer: Berlin/Heidelberg, Germany.

Tsukerman. (2019). Machine learning for cybersecurity cookbook : over 80 recipes on how to implement machine learning algorithms for building security systems using python (1st edition).

Packt Publishing, Limited.

TraceSecurity, (2022) Mitigating IT security risk with penetration tests, TraceSecurity,

Whitepaper. [Taking a Risk-Based Approach to PTing (isaca.org)](https://www.isaca.org/resources/news-and-trends/industry-news/2022/taking-a-risk-based-approach-to-pen-testing#Converting%20Pen%20Test%20Results%20to%20Information%20Security%20Risk)

Wack, J., Tracy, M., and Souppaya, M., (2003). NIST special publication 800-42, guideline on network security testing. Computer Security Division, National Institute of Standards and Technology, pages 1–92.

Whitehouse, T., (20 May 2022), 3 ways to transform risk assessment with ai, Wall Street Journal, Risk & Compliance Journal, <https://deloitte.wsj.com/articles/3-ways-to-transform-risk-assessments-with-ai-01652994728?tesla=y&tesla=y>