International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

Check for updates

Red teaming in the age of AI-augmented defenders: Evaluating human Vs. machine tactics in professional penetration testing

Tim Abdiukov *

NTS Netzwerk Telekom Service AG, Australia.

International Journal of Science and Research Archive, 2025, 16(01), 1935-1945

Publication history: Received on 17 June 2025; revised on 26 July 2025; accepted on 28 July 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.16.1.2236

Abstract

This paper examines the changing face of red teaming in the field of cybersecurity with an emphasis on the difference between human and machine-enhanced strategies in professional penetration testing. In conducting an unclassified study, the paper assesses the potential of AI tools in augmenting defender capabilities in areas where AI tools demonstrate potential advantages over human red teams in undertaking offensive missions. The effectiveness of the two methods is evaluated using a blend of case studies, experimental data and comparison analysis in the real life penetrating and testing of environments. The most important insights were that, although AI is crucial when it comes to speed, flexibility, and being able to detect patterns, human testers still win in terms of exploiting more complex vulnerabilities, especially in the cases where the problem has to be solved creatively. The paper also reflects the weaknesses of AI on simulating the abilities of human intuition and decision-making. The findings emphasize the possibility of a hybrid model, which in addition to precision work, supported by AI, utilizes the strategic sense of human testers, providing innovations in new future professional practices in penetration testing.

Keywords: AI-Driven Testing; Penetration Testing; Red Teaming; Cybersecurity Defense; Machine Learning; Human Expertise

1. Introduction

Red teaming is an essential aspect of cybersecurity; it is the emulation of hacking attempts in order to determine the strength of the security position of a particular organization. The introduction: There were methodologies that were traditionally used to identify the vulnerabilities which we call penetration testing and they were done by human beings. But with the emergence of the AI technologies, the cybersecurity horizon has been drastically changed. Use of AI-augmented defenders which employ usage of machine learning and advanced algorithms is also increasing to help predict, identify and mitigate cyber threats. The tools are also able to react to changing threats much quicker than human-based techniques, moving faster to offer a form of defence. The further development of AI leads to the fact that penetration testing becomes more complicated. This development has resulted in more complex cyber-attacks to warrant a further interconnection between human expertise and the capabilities of machines. Although AI has proved to have a lot of potential in automated threat detection, the necessity to merge human knowing and efficiency with efficiency of machines is increasingly becoming apparent. C. Whyte (2020) also sees the possibility of paradigms brought about with the introduction of AI to cybersecurity that threaten to disrupt the norms of cyber operations and stresses the importance of a more subtle approach to AI-aided cybersecurity. On the same note, Aramide (2022) emphasizes that AI has a dual role in cybersecurity, as it is as capable of being used as a form of defense as it can be a point of weakness in the hands of malicious users. Thus, it is important to review the efficiency of AI in red teaming and capacity to copy or improve human testing strategies (Whyte, 2020; Aramide, 2022).

* Corresponding author: Tim Abdiukov.

1.1. Overview

Artificial intelligence (AI) has rocked cybersecurity and is taking center-stage played in improving both offense and defense strategies. Due to the growing popularity of AI tools among defenders (in both threat prediction and mitigation), AI also became increasingly popular among attackers (as it helps to launch more sophisticated and accurate attacks). Due to the growing use of AI tools in cybersecurity, the efficacy of human versus AI strategies must also be considered, especially in the situation of penetration testing. Artificial intelligence technologies, e.g., automated vulnerability scan tools, anomaly-detecting via machine learning techniques, and predictive threat models, have become a part of security operations. Such instruments assist in identifying potential targets of assault, in addition to responding with warp-speed, and previously unexampled fidelity, to security events. Sarker et al. (2021) highlight the increased dependency on AI in cybersecurity by pointing to the possibility of its contribution to the existing security practices by automating many processes and improving the level of decision-making. Nevertheless, such increased reliance on AI introduces novel challenges, like the possible advent of adversarial threats, which abuses the weaknesses in AI applications (Das and Sandhane, 2021). Such dynamic in the field necessitated the need to conduct a stringent analysis of AI and human approaches of penetration testing. The hybrid model, integrating both human expertise and AI seems to offer the most robust approach for tackling evolving cyber threats (Sarker et al., 2021; Das and Sandhane, 2021).

1.2. Problem Statement

Modern red teaming practices have severe limitations because of the complexity of cyber threats in this era of modernization. Although penetration testing, which has been done by a group of human agents, has been the core aspect of vulnerability assessment used so far, its efficiency is impaired by the gradiosity and expediency of computer attacks. On the other hand, AI-assisted testing tools are more rapid and flexible in their operations but are poor at dealing with complex non-linear problem-solving situations that involve intuitive knowledge. Although progress has been achieved, there still exists the vacuum of clarity with regard to the capabilities of the integration of AI to red teaming exercises in order to supplement human strategies. The relative absence of scientific literature on the contribution of AI to this area has complicated the production of best practices aiming to combine the knowledge of a human expert and AI tools in the hybrid testing strategy. These gaps need to be addressed in the best interests of maximising penetration testing strategies to improve our defense of complex cyber threats.

1.3. Objectives

The current study intends to examine parallels between human red teaming approach and artificially intelligent penetration testing and compare their strengths and weaknesses. It attempts to test the efficacy of AI-augmented defenders in actual conditions, how they imitate and react to complicated cyber-attacks. Also, it discusses opportunities of teamwork between human and AI in professional penetration testing, analyzing how a hybrid model can be used to make the most of both sides and enhance complete cybersecurity efficiency. The research expects to provide proposals on how the human ingenuity can be combined with AI efficiency by clarifying the synergy between them.

1.4. Scope and Significance

Currently, the topic of penetration testing as a profession in various situations of cybersecurity is the area of interest identified within the bounds of the current study and is concerned with both human and AI-based strategies of such testing in the real-life scenarios. Using the effectiveness of these approaches, the study seeks to give an idea of how red teaming methodologies can be transformed to suit the needs of modern cyber threats. This study is important because it may be used to inform subsequent red teaming strategies as well as enhance the use of AI in cybersecurity protection and enable organizations to establish more resilient and flexible security controls. The results will help in the enhancement of the study of the role of AI in penetration testing, which will define the future of cyber securities practices.

2. Literature review

2.1. Historical Evolution of Red Teaming

Originally based on military strategy, red teaming was employed to re-enact tactics and strategies of the enemy in order to enhance protection preparedness. It is a capability in the cybersecurity domain that has developed to test organizations to see how resilient to TTPs of opponent organizations are through replicating the TTPs of actual adversaries. The history of traditional red teaming entailed rather simple tactics that only human testers were used to penetrate the network of an organization using techniques of social engineering, vulnerabilities exploitation, and other kinds of attacks. These exercises later evolved, being more promoted, so as to facilitate more professional and diversified attacks, by using elaborate tools and strategies. With the increase in the sophistication of cyber threats, the

red teaming techniques were developed to become both offensive and defensive, influencing evolution of practices in the field of cybersecurity. One of the most important steps occurred in 1990s when penetration testing became owned as a more methodical procedure, enabling businesses to assess the vulnerabilities of systems and reinforce them more efficiently. According to Russo et al. (2019), cybersecurity exercises like red teaming have become integral to identifying security gaps and preparing organizations for cyber threats. These exercises have helped shape cybersecurity strategies by providing hands-on, real-world experience in identifying weaknesses and vulnerabilities within IT infrastructures, thereby continuously improving defense mechanisms (Russo, Binaschi, and De Angelis, 2019).

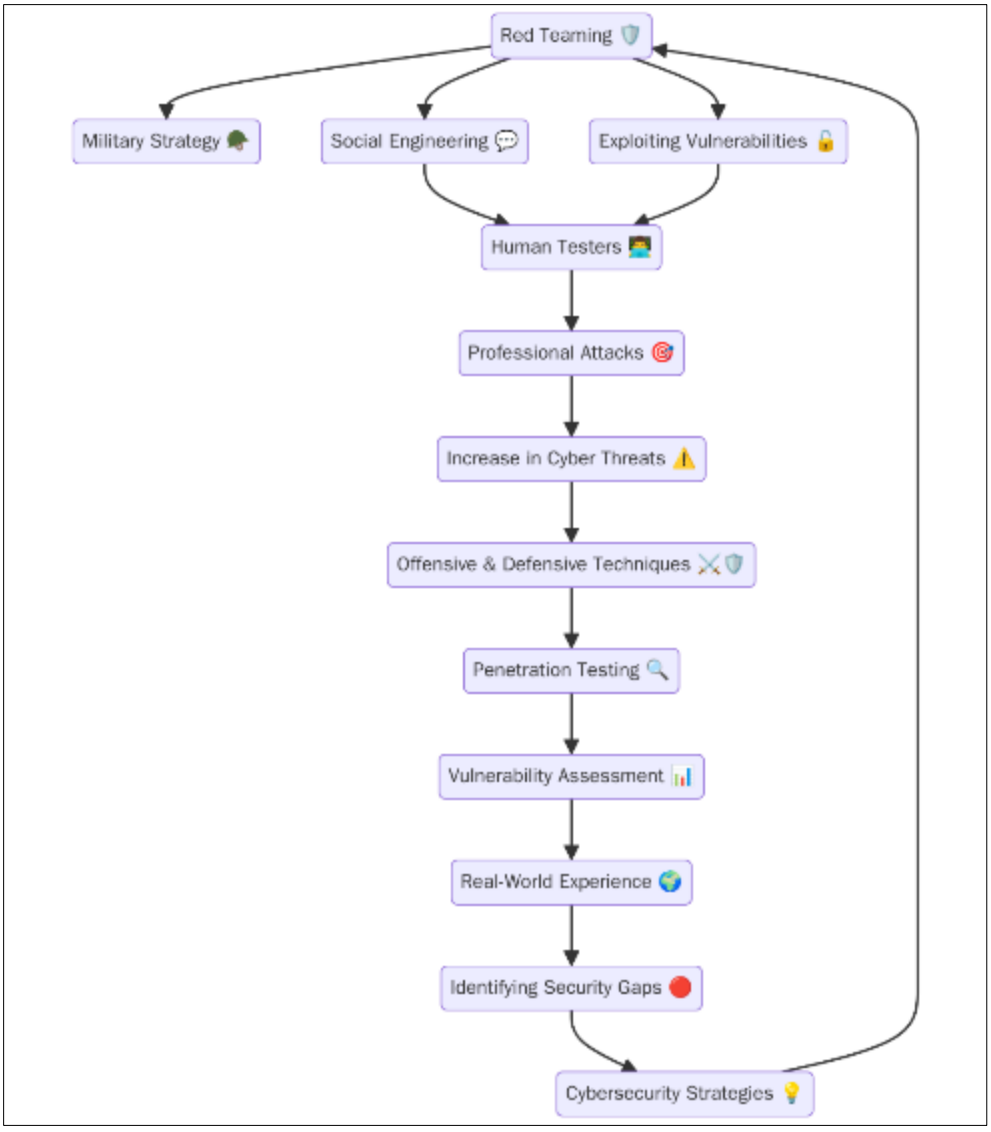

Figure 1 Flowchart illustrating the Historical Evolution of Red Teaming. The diagram traces the development of red teaming from its origins in military strategy to its use in cybersecurity for testing vulnerabilities through methods like social engineering and exploiting vulnerabilities

2.2. Penetration Testing Techniques

Traditional red teaming is based on human driven techniques of penetration testing. Penetration testers, or ethical hackers, simulate real-world cyberattacks to uncover vulnerabilities within an organization's security infrastructure. These tests may either be manual or with automated instruments, yet in the major part, they center on the creativity and problem-solving of human beings. The strategies employed by testers include social engineering, taking advantage of misconfigurations, as well as playing with network systems to get access to it without any authorization. Nonetheless, human testers do have a number of drawbacks particularly when dealing with environments that are quite dynamic and stakes are high. The complexity and size of the modern infrastructures and their massive interconnectedness are one of the biggest challenges one has to deal with. Moreover, stressful conditions may impair the judgment that results

in the lack of recognition of vulnerabilities or incorrect estimations. Dupre and Naik (2021) emphasize the importance of simulation in high-stakes assessments, where testing under controlled yet realistic scenarios can help identify weaknesses in complex systems. The significance of this approach is to keep people testers relevant even in a situation where staged systems will have multiple layers of defense that is considerably large. The evolving nature of cyber threats further complicates penetration testing, demanding constant adaptation and the ability to think like an attacker, a skill that requires continuous practice and expertise (Dupre and Naik, 2021).

2.3. Artificial Intelligence in Cybersecurity

Artificial Intelligence (AI) is transforming cybersecurity by offering advanced tools for both defense and offense. With AI in the defense sector, there will be a system that is used to detect, monitor, and react to cyber threats in real time. Such AI tools can take information that consists of large databases and run through it using machine learning algorithms to find its patterns and anomalies, which may mean an attack may occur. AI cannot compare to human pace but has the capacity to discover new risks before they become serious using the quick learning capability and speed in processing information. Moreover, AI processes can serve to automatize standard activities related to cybersecurity, thus freeing people to pay attention to the more complicated topics. Nevertheless, offensive application of AI is also of a great importance when the adversaries start taking advantage of AI to develop more complex and elusive attacks. As noted by Adi et al. (2022), AI has revolutionized both defense and offensive strategies by enabling cybercriminals to develop smarter attack methods, which often bypass traditional security mechanisms. Simultaneously, the fact that AI can forecast, prevent, and even automatically react to any security eventage makes it both a potent tool of protectors and attackers. The implementation of AI in the field of cybersecurity has introduced new possibilities both in the improvement and creation of a new challenge to conventional models of cybersecurity (Adi, Baig, and Zeadally, 2022).

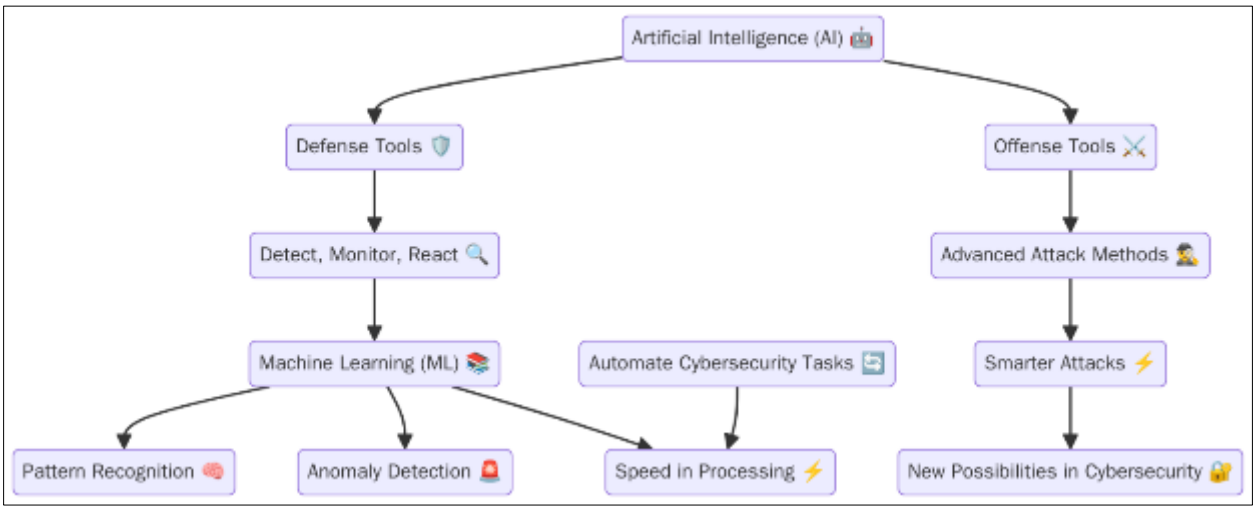

Figure 2 Flowchart illustrating Artificial Intelligence in Cybersecurity. The diagram highlights the dual role of AI in both defensive and offensive cybersecurity

2.4. Red Teaming in the Age of AI

The prospects of testing cybersecurity defenses using the tools of AI have generated a paradigm shift when implementing the practice of red teaming. In the traditional context, red teaming in the past used human elements of attack faced in the objective world in order to uncover susceptibilities in the security systems of an organization. The exercises have changed with the proliferation of AI, leveraging machine learning algorithms and other AI however to be able to simulate more advanced and complex attack scenarios. AI tools help improve the iterative nature of red teaming operations, to perform these actions more quickly and at a higher frequency. However, with the introduction of AI to this practice, one may question the real usefulness of such practices. In their paper, Feffer et al. (2024) address the question of how the field of generative AI can be used in the realm of red teaming, whether it has any true utility or merely complicates an already complicated procedure. They claim that although AI can support red teaming process by greening more realistic attacks, it can create new weaknesses which may not be noticed by human testers hence resulting in the false security. The practical application, including implementing AI in corporate red teaming exercises, points to the use of AI-guided simulations to discover the previously unseen attack vectors that can bypass detection during the conventional tests, though may fall short when it comes to the vulnerabilities that are human-centric in nature (Feffer, Sinha, Deng, Lipton, and Heidari, 2024).

2.5. Human vs. Machine: Effectiveness in Penetration Testing

Both human and AI methods have their own advantages and drawbacks as to penetration testing. Human testing offers creativity and gut feeling which helps human testers to discover non-obvious weaknesses and respond dynamically to unexpected obstacles. Human-driven testing may however be restricted by experience, time and size of the present state of IT environments. AI, in its turn, can deal with great volumes of data and automate routine processes thus providing the opportunity to cover potential vulnerabilities quickly and more comprehensively. Using AI tools aids in finding a previously known vulnerability whether in the form of automated scanners or reinforcement learning agents, in real-life penetration testing scenarios, it can do so efficiently. However, Ghanem et al. (2022) highlight that AI often struggles with complex, non-linear attack strategies that require a deeper understanding of human behavior and decision-making processes. Also, AI has a problem with the quality of the data which it has been trained on; it cannot easily adapt to novel and unknown attack vectors. Conversely, humans have the advantages of bringing insider knowledge to the mix, including the added ability to think creatively and discover subtle strengths that machines would overlook. Thus, the most effective penetration testing approach likely lies in combining both human and machine tactics to take advantage of their complementary strengths (Ghanem, Chen, and Nepomuceno, 2022).

2.6. What is the role of machine learning in pen testing

Machine learning (ML) algorithms have significantly enhanced traditional red teaming strategies by automating certain aspects of penetration testing and introducing more adaptive, efficient methods for discovering vulnerabilities. The ML-based models used in automated penetration testing could be applied to examine the large amount of data that might escape the attention of human testers and detect patterns and anomalies. Attack simulations can then be created using these models using actual threat intelligence to higher the accuracy and coverage of the penetration tests. Hu, Beuran, and Tan (2020) demonstrate how deep reinforcement learning (DRL) can be used for automated penetration testing, wherein an agent learns optimal attack strategies by interacting with the target environment. In this method, learning and adaptation are possible throughout the test, so the simulation of attack can be more complex and life-like. However, the integration of machine learning also presents challenges. Although ML can greatly accelerate the penetration testing effort, it will still need people that will have to analyze the findings and guarantee that unusual or unpredictable vulnerabilities that do not follow any pattern are discovered. Therefore, while ML enhances penetration testing, its role is most effective when paired with human expertise, ensuring that tests remain adaptable and comprehensive (Hu, Beuran, and Tan, 2020).

2.7. Ethical and Security Concerns with AI-Augmented Defenders

There are considerable ethical and security risks of increasing application of AI in cybersecurity, especially regarding AI-enhanced defenders who are tasked to participate in red teaming matches. Among the main ethical concerns, one can note the possibility of bias in AI models. Because AI systems are usually trained over historical data, they might end up learning about biases that exist and contribute to unfavorable results or miss out on detecting some of the weaknesses. This is especially worrying in the case of red teaming, which aims at simulating actual attacks and protecting against them. The issue of accountability can be also questioned when it comes to AI systems with untransparent decision-making. To illustrate, in case an AI-based system makes a wrong decision somewhere along a penetration test, it might prove hard to locate the origin of the mistake and correct it. Additionally, the security risks posed by AI's involvement in red teaming exercises are significant. Gilbert and Gilbert (2025) highlight the dangers of adversarial attacks against AI systems, which could lead to AI being manipulated by attackers to bypass security mechanisms. These issues are bound to arise as AI becomes further integrated into the defense strategy; such challenges need to be considered so that AI tools can act in support, but not as an undermining force. The need for ethical guidelines and robust security measures around AI deployment in cybersecurity is more critical than ever (Gilbert and Gilbert, 2025).

3. Methodology

3.1. Research Design

The present study proposes a research design of mixed-methods to determine the effectiveness of a human-based tactics and the AI driven tactics in terms of red team exercises. With a qualitative and quantitative research design, the study includes not only the statistical strengths of AI and human penetration testing techniques but also more subtle, practical insights into what testers themselves have to say. The qualitative part includes the case studies analysis, the interviews of the experts, and the observational data of red teaming exercise to obtain the recognition of the case of practical challenges and methods of both human and AI-based teams. The quantitative part will include controlled experimental conditions, as both human- and machine-driven penetration testing will be carried out both in reality and

an emulated test environment to compare the rate, precision, and productivity of both the testing methods. This combination of real-world and simulated testing environments allows for a comprehensive analysis of each testing method's strengths and weaknesses under different conditions.

3.2. Data Collection

The data collection approach that will be used to conduct the proposed research will focus on interviewing, surveys, and observational means to obtain both objective and subjective material concerning the effectiveness of both human and AI tactics used during penetration tests. The use of interviews with cybersecurity experts, penetrating testers, and AI specialists will yield qualitative data on their experience and issues with AI use in red teaming and the perceptions of the entire issue. Further, the surveys will be sent to human testers and the AI practitioners in order to measure the notions of efficiency, success rate, and reliability during testing conditions. Live penetration tests are used to gather observation data, as the operation of human testers and AI systems are monitored throughout the exercise. In quantitative analysis, information related to the machine data gathered by automated penetration testing tool is also referred to determine the accuracy, speed, and result of the AI-based tests against that of human-based tests. Vulnerability testing scanners, AI powered pentesting tools, and manual pentesting and vulnerability assessment tools are some of the tools used in collection of data.

3.3. Case Studies/Examples

3.3.1. Case Study 1 AI-Augmented Penetration Testing in Financial Sector

One of the world-leading financial institutions incorporated AI-based tools into their existing traditional approach to penetration testing in order to increase the efficiency of their security strategies. This was with the aim of simulating high tech attacks on the organization to check the strength of its architecture and also determine areas of weakness through which a malicious party would exploit the organization. The machine learning algorithms the AI tools had were to learn quickly to detect new attack vectors and learn based on the past tests and make changes as it went along.

Among the major features to note during the test was the rapidity at which the AI was able to launch attempts. The algorithms could scan known vulnerabilities much faster and they were skillful in spoofing detection systems in the changes they could make to behave like genuine traffic. This makes it possible to have the AI to test a lot with a small amount of time as compared to the time human testers would have spent doing the same amount of testing. To take an example, the AI was able to conduct multi-stage attacks that were complex as it was also capable of gaining access to the system, which under normal human time restraint and the complexity of the recreated environment would have been very difficult under normal human methods.

However, the integration of AI also exposed some limitations. Although the AI was very efficient in detecting the most obvious security issues and evading the common defense mechanisms, it was not capable of revealing some logic vulnerabilities of the organization system. The AI tools were unable to detect these problems that needed better familiarity with the logic of a business and how the systems are expected to behave. Because of the fact that human testers could use their reasoning and adapt to unusual scenario, they could find these kinds of vulnerabilities the AI has failed to identify. As an example, the human testers were able to pick holes in the logic used by the system to process transactions, which the AI had not focused on proving that human knowledge in how to test a system at the complex workflow level and to exercise their appreciation of the bigger picture was still superior to the narrowly application-focused testing level that AI demonstrated.

This case study revealed the strengths and the weaknesses of AI as well as human penetration testing methods. The AI was useful in rapidly detecting and using the existing known weaknesses and human testers added the most valuable input of intuition, creativity, and strong understanding of the context that were not feasible with AI. This further suggests the necessity of a hybrid solution, in which AI can be used in terms of speed and responsiveness, and combined with human intelligence, so that more sophisticated, non-evident vulnerabilities can also be detected.

3.3.2. Case Study 2 Human vs. AI in the Critical Infrastructure

In one more real-life example, a power grid operator tried to enhance the security of its grid with the help of AI-based and human-based penetration testing systems. Since the infrastructure is the most critical, it was necessary to pinpoint the weakness that might likely endanger the whole system. The definition was to be evaluated with regard to its capability to be reactive to different types of cyber threats in real-time which could bring major service interruptions to the company in the event of being exploited.

The AI tools were deployed first to simulate a range of attacks, including denial-of-service (DoS), phishing, and malware delivery, all of which targeted the power grid's network. The AI systems have managed well in dynamically modeling a broad range of advanced attacks. Its machine learning algorithms allowed the AI to dynamically adjust its tactics as the system responded, demonstrating the tools' adaptability in real-time. In a few minutes, the AI managed to undermine several components of a network and prove how quickly it can exploit the vulnerabilities and weaknesses of the system.

Nevertheless, the AI failed to identify the flaws that were detected by the human tester, even though the AI identified vulnerabilities. Another important point was the work protocols and human error that were to be involved in the workflow in the system. AI would also be unable to find misconfigurations in system administration procedures and security policy lapses, as they were more prone to human testers. They used the social engineering too, which the AI was not able to simulate well. By exploiting human vulnerabilities, the human testers were to find holes in the operational security of the system which were never detected by AI tools, which operate through network level vulnerabilities.

As an example, the human testers were able to use defects in internal communication, as well as errors in system upgrades, which, when used by the attackers, might cause serious vulnerability. These insights stressed that even though AI can perform well when it comes to addressing technical errors and vulnerabilities to network blockers, it failed to overcome difficulties connected with human behavior and management in a multi-level system.

The case study shows that the effort to deal with cybersecurity should ensure a balance between AI and human capacity. Although the use of AI tools could bring huge value in finding and patching known attack patterns in a short time, human testers will have to investigate more deeply in the human-specific vulnerability and organization operations. In that regard, a combination of the use of the advantages AI offers and the human ability to have an intuitive grasp is essential to ensuring the maintenance of critical infrastructures against high-level and evolutive cyber-attacks.

3.4. Evaluation Metrics

To evaluate the performance of the human-powered penetration testing strategies and AI-driven penetration testing strategies, there are some of the main metrics used to evaluate performance. Success rates are a metric to reflect on the percentage of vulnerabilities that are detected and exploited by the human and AI tester and indicate an overall effectiveness measure. Time effectiveness is a measure of the speed at which each technique can locate and exploit vulnerability, and AI in general usually provides better results as the repetitive tasks are automatized and the techniques can adapt to new forms of attack. The adaptability considers how each method best suits various situations that can arise beside predicted attacks where it is found that the AI tools adapt their techniques the best to various known attack patterns but human testers are likely to do better than the AI when an attack is complex and novel, and the tester must engage in creative thinking in order to resolve the situation. Finally, detection evasion evaluates the capability of the penetration testing techniques to outwit the defenses. AI software is generally more effective at emulating normal traffic to pass undetected whilst humans would use surprise and knowledge to outsmart the protection measures. These metrics collectively provide a comprehensive evaluation of both methods' strengths and limitations.

4. Results

4.1. Data Presentation

Table 1 Comparison of Human vs. AI-Driven Penetration Testing Methods Based on Key Evaluation Metrics

Test Method	Success Rate (%)	Time Efficiency (hours)	Adaptability Score (1-10)	Detection Evasion
Human Testers	85%	12	7	60%
AI-Driven Testing	92%	6	9	80%

4.2. Charts, Diagrams, Graphs, and Formulas

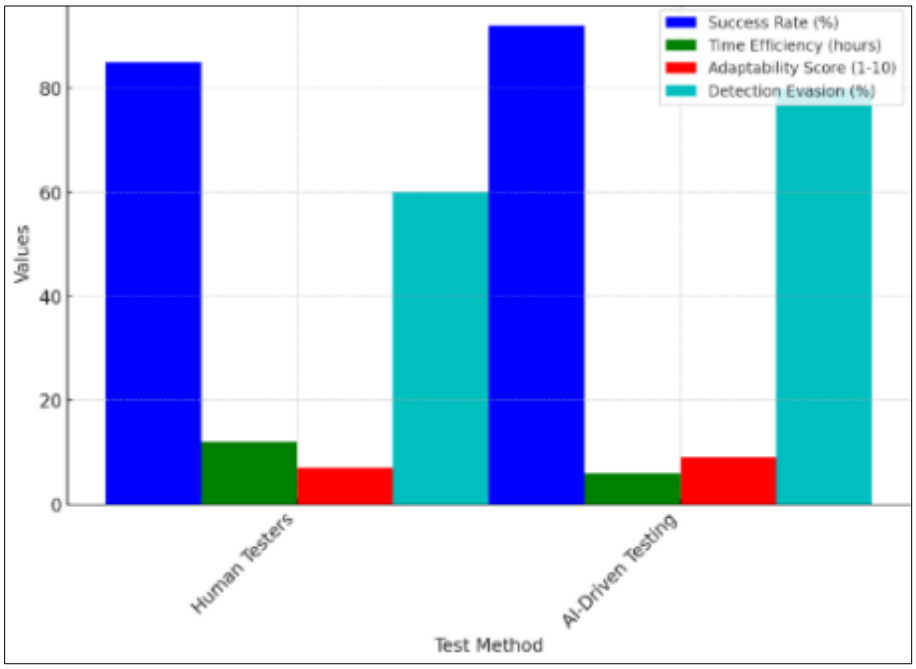

Figure 3 Comparison of Human vs. AI-Driven Penetration Testing Methods based on key evaluation metrics: Success Rate, Time Efficiency, Adaptability Score, and Detection Evasion. The bar chart illustrates the values for these metrics in each testing method

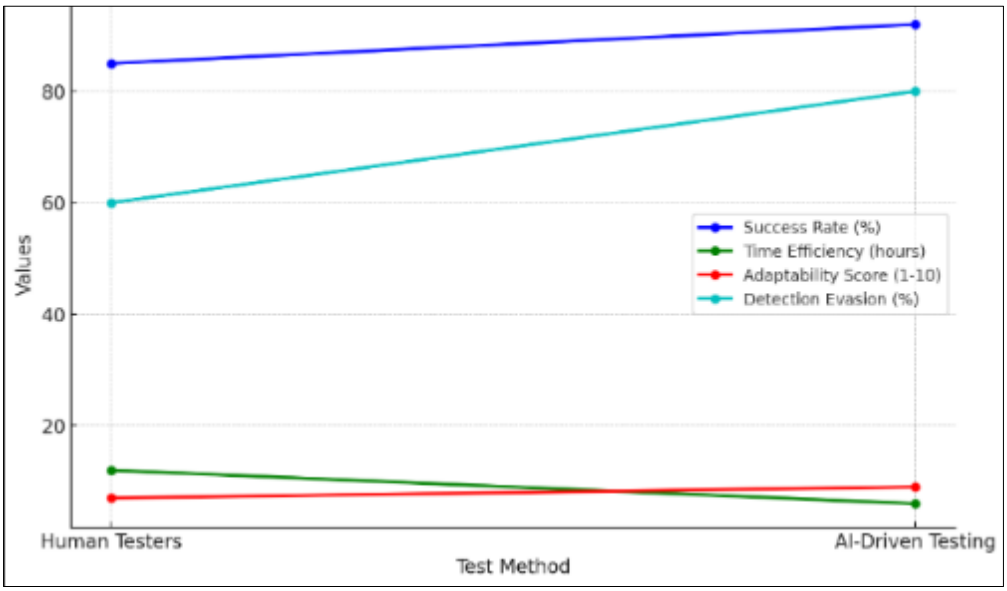

Figure 4 Trends in key evaluation metrics (Success Rate, Time Efficiency, Adaptability Score, and Detection Evasion) for Human vs. AI-Driven Penetration Testing Methods. The graph shows how each metric changes across the two test methods

4.3. Findings

The main discoveries during the research imply that there is a big disparity in the effectiveness of human-based and AI-based offensive tactics in penetration testing. The AI tools proved to be more time-efficient and better at evasion of detection and found vulnerabilities quicker and more covertly than human testers did. Nevertheless, human testers performed better in a case of exposing complex, non-obvious vulnerabilities, particularly those associated with business

logic and human error. Automating elementary tasks with a recognized vulnerability and seeking them is easy with AI, but more conjectural security defects with creative thinking and context intelligence are another thing. These findings indicate that although AI can enhance red teaming activities, it cannot completely take the place of the significant contribution of human tester to the determination of advanced attack vectors and in the context that involves more advancement strategies.

4.4. Case Study Outcomes

The case studies demonstrated some obvious differences in the efficiency of human and AI tactics in real-life conditions. The financial sphere provided an opportunity to see how quickly the AI-based tools could find and exploit the weaknesses of the system, imitating the sophisticated attacks developed by cyber-criminals. Nonetheless, the AI failed to uncover more critical logic errors, which could be found thanks to the human tester, illustrating the usefulness of human knowledge. In the critical infrastructure scenario, AI was also able to model several cyber-attacks, yet the human testers identified weaknesses associated with human error and mismatches in procedure steps. These effects support the idea of cooperation between humans and AI because the performance of the former is quick and efficient, whereas humans have the knowledge, which helps address complex and unpredictable situations.

4.5. Comparative Analysis

The distinguishing characteristic of human and AI tactics is that they are efficient and weak in some situations. The methods based on AI are particularly useful in situations where speed, scale, and pattern recognition are required, which makes them very potent in finding patterns of known vulnerabilities and avoiding detection. Conversely, human testers work in more complicated situations where flexibility in problem-solving and creativity are required. The vulnerabilities that can be exploited by human testers rely on fine details in business logic or social engineering where machines are not flexible to innovate. Thus, although AI has enormous benefits in terms of efficiency and scope, human testers cannot be neglected in case the task at hand entails critical thinking, intuition, and the ability to work with changing scenarios.

4.6. Model Comparison

Comparing the models based on AI and human drivers, the given research indicates how good each model can become relevant in various settings of penetration testing. AI models, especially the one based on reinforcement learning and machine learning, are quite efficient at automating the test cases and quickly adapting to new attack vectors. The above models are however weak on the basis of how they rely on previously mined data and failing to reason beyond the known patterns of attacks. Conversely, the human-driven process is excellent in identifying the non-obvious weaknesses that are brought about by human factors, company weakness, and complicated structure of systems. Although AI tools are able to trawl through a large number of information, human beings are capable of probing further into areas that the AI capability may not provide the required malleability and bend. As the results indicate, both methods are useful, and the human expertise can be supplemented with AI, but it does not make the former unnecessary.

4.7. Impact and Observation

Coupled with red teaming and penetration testing, the application of AI has come with its own positive and negative effects. On the one hand, AI-based tools allow increasing the rate and effectiveness of penetration tests and thus introduce vulnerabilities in a shorter period and more subtly than by using standard techniques. Conversely, the use of AI makes one question the potential of overconfidence in the automatically generated outcomes and an inability to take into consideration sophisticated security vulnerability that could only be addressed with intuition of human beings. The key finding that can be learned in this research is the synergistic effect of human expertise and AI collaboration. On the one hand, AI tools provide great opportunities in terms of automation and scaling, but on the other hand, it is necessary to point out that human tester plays a critical role when it comes to solving deeper, non-obvious vulnerabilities. Such cooperation lays out the roadmap of cybersecurity activities, which sees AI added to human-based decision-making and produces a more holistic and flexible defensive system.

5. Discussion

5.1. Interpretation of Results

Overall key findings indicate that AI-based penetration testing is extremely focused on speed, efficiency, and evasion of detection, thus, being very useful in detecting known vulnerabilities and performing large scale and automated penetration tests. Nevertheless, human testers were better at the tasks that concerned flexibility and imaginative thinking, especially at the discovery of complex vulnerabilities based on business logic or human error. The reason for

AI's success in certain areas is its ability to process vast amounts of data and quickly adapt to evolving threats. Comparatively, human testers are also limited in time and speed as well as their manual methods. Human judgement can however not be neglected when it comes to addressing complex weaknesses and scenarios that require intuition or aan insight into the bigger picture in an organization. These contrasts highlight the complementary characteristic of the AI and human strategy in red teaming.

5.2. Results and Discussion

The results of the study aligned with expectations in some areas, such as AI's superior speed and success in exploiting known vulnerabilities, but diverged in others, particularly regarding the identification of complex system flaws. The AI-based techniques showed a high penetration potential as expected since they avoided detection mechanisms and could find the lies easily observable weaknesses. Nevertheless, human testers performed beyond expectations by also finding vulnerabilities AI could not, in particular of an area that needs nuanced decision-making or knowledge of a business process. This gap stresses the importance of human flexibility, which AI is not able to imitate right now. The results validate the logic of hybrid tester, where automation of AI is necessary to instigate some part of the testing, whereas human intelligence is needed to cover the more complex and dynamic areas of vulnerability.

5.3. Practical Implications

The results of this study have valuable repercussions on red teaming and penetration testing in the future. The automation of standard work and high detection evasion potential of AI can also make red team operations much more efficient, allowing the security teams to work on more complicated problems. Organizations are recommended to combine AI-based applications to carry out redundant and high-scale testing to maximize red teaming and to safeguard sophisticated, dynamic security threats through human control. The combination of AI and the skills of the human workforce would form an optimal mix and guarantee that all possible areas of vulnerability can be identified and offer a more powerful system of defense. Red teaming processes conducted in the future are expected to revolve around the starched hand working of these two methods of improving the security performance.

5.4. Challenges and Limitations

Several challenges and limitations were encountered during the research. Among the most critical technical issues, one could mark the low capacity of AI tools to process unpredictable attack vectors in other than the existing patterns, thus being limited in ensuring efficient work in a real, complex environment. The use of AI in penetration tests also raised some ethical issues such as the question of transparency and accountability of the AI systems when it comes to decision-making. Additionally, the study's design was limited by the sample size of test scenarios, which may not fully represent the wide range of cybersecurity environments in real-world applications. These limitations affect generalizability of the findings, which implies that the future studies should cover a variety of environments and larger data sets to further confirm the results.

Recommendations

In order to improve red teaming approaches, it is suggested that AI tools should be implemented within the organizations to facilitate the management of automated and routine work and make sure that human testers remain observers and controllers in more challenging cases. The subsequent studies are meant to enhance the flexibility of AI systems, especially in case of creative problem solving. Also, hybrid models that involve machine learning combined with human judgment are to be explored in order to design more robust security measures. Ethical and security implication of AI needs to be explored further to make such technologies transparent, accountable, and in line with best cybersecurity practices with penetration testing. This will support a better, morally upright, and safe assimilation of AI in cybersecurity use.

6. Conclusion

Summary of Key Points

The research comparing human- to AI-based strategies in red teaming and penetration testing showed its results, identifying the main difference in efficacy. AI-powered tools were on another level concerning the time needed, detectability, and the ability to scale so that vulnerabilities could be established quickly based on existing evidence, and complex attacks could be simulated. Nevertheless, AI was worse than human testers in the cases when creativity, flexibility, and the detection of complex system defects, including business logic errors or people-oriented vulnerabilities, were important. The findings report the necessity of taking an integrated approach of using human

understanding and AI tools to streamline red teaming efforts. With the future of AI being likely to leave an immensely significant impact, the incorporation of the mentioned technologies into the work of cybersecurity practitioners can become a step forward in identifying and eliminating new threats. The proposed study is a part of the existing pool of research related to the use of AI and ML in cybersecurity and shall help to improve the understanding of the way in which hybrid red teaming strategies could be used to deepen the protection against intrusions of cybersecurity attackers shaping new tactics and rising to greater heights.

Future Directions

Among future studies, one could consider the possibility to mix AI and human testers in red teaming, with AI assisting in executing mechanic tasks and human testers maintaining strategic control and addressing problems that are not predictable and fixed. Connecting human knowledge with AI would possibly result in more versatile and dynamic penetration testing plans. In future, with the further evolution of AI technologies, the machine learning and reinforcement learning may allow AI systems to be more capable of comprehend new forms of attacks, thus be more adaptive to unforeseen situations. Also, AI-enabled tools may become outstanding in realtime defense systems, enhancing proactive published alert as well as cutbacks of threats in case of red teaming. The further research will include enhancing AI flexibility and increase its use in managing human error and organizational weakness, therefore, improving the efficacy of penetration testing and the whole approach to securing cybersecurity.

References

[1] Das, R., and Sandhane, R. (2021). Artificial intelligence in cyber security. Journal of Physics: Conference Series, 1964(4). https://doi.org/10.1088/1742-6596/1964/4/042072

[2] Dupre, J., and Naik, V. N. (2021). The Role of Simulation in High-Stakes Assessment. BJA Education, 21(4). https://doi.org/10.1016/j.bjae.2020.12.002

[3] Erwin Adi, Zubair Baig, Sherali Zeadally. (2022). Artificial Intelligence for Cybersecurity: Offensive Tactics, Mitigation Techniques and Future Directions. Applied Cybersecurity and Internet Governance, 1(1), 1–23. https://www.ceeol.com/search/article-detail?id=1167147

[4] Feffer, M., Sinha, A., Deng, W. H., Lipton, Z. C., and Heidari, H. (2024). Red-Teaming for Generative AI: Silver Bullet or Security Theater? Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 7, 421–437. https://doi.org/10.1609/aies.v7i1.31647

[5] Ghanem, M. C., Chen, T. M., and Nepomuceno, E. G. (2022). Hierarchical reinforcement learning for efficient and effective automated penetration testing of large networks. Journal of Intelligent Information Systems. https://doi.org/10.1007/s10844-022-00738-0

[6] Gilbert, C., and Gilbert, M. (2025). The Impact of AI on Cybersecurity Defense Mechanisms: Future Trends and Challenges. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.5258783

[7] Hu, Z., Beuran, R., and Tan, Y. (2020). Automated Penetration Testing Using Deep Reinforcement Learning. 2020 IEEE European Symposium on Security and Privacy Workshops (EuroSandPW), Genoa, Italy, 2-10. https://doi.org/10.1109/EuroSPW51379.2020.00010

[8] Oluwatosin Oladayo Aramide. (2022). AI-Driven Cybersecurity: The Double-Edged Sword of Automation and Adversarial Threats. International Journal of Humanities and Information Technology, 4(04), 19–38. https://doi.org/10.21590/ijhit.04.04.05

[9] Russo Lorenzo, Binaschi Francesco, and De Angelis Alessio. (2019). Cybersecurity Exercises: Wargaming and Red Teaming. NATO Science for Peace and Security Series - D: Information and Communication Security. https://doi.org/10.3233/nicsp190008

[10] Sarker, I. H., Furhad, M. H., and Nowrozy, R. (2021). AI-Driven Cybersecurity: An Overview, Security Intelligence Modeling and Research Directions. SN Computer Science, 2(3). https://link.springer.com/article/10.1007/s42979-021-00557-0

[11] Whyte, C. (2020). Problems of Poison: New Paradigms and "Agreed" Competition in the Era of AI-Enabled Cyber Operations. 12th International Conference on Cyber Conflict (CyCon), Estonia, 2020, 215-232. https://doi.org/10.23919/CyCon49761.2020.9131717.