jamk

Artificial Intelligence in Red Teaming

Mays Al-Azzawi

Bachelor's thesis
December 2024
Bachelor of Engineering, Information and Communication Technology

jamk | Jyväskylän ammattikorkeakoulu
University of Applied Sciences

jamk

Description

Al-Azzawi Mays

Artificial Intelligence in Red Teaming

Jyväskylä: Jamk University of Applied Sciences, December 2024, 24 pages (+ appendices).

Bachelor’s degree in information and communications technology

Permission for open access publication: Yes

Language of publication: English

Abstract

Red teaming involves simulating real world attacks on targets such as organizations, infrastructure, or individuals to test their defences and assess the vulnerabilities. Artificial intelligence plays a significant role in red teaming cyberattacks. The thesis explores the impact of AI in red teaming by examining how AI methods can be misused in various scenarios and identifying the typical targets for these attacks. Recent studies highlight the risks associated with large language models (LLMs), a form of advanced AI, and their potentials to reshape the red teaming domain. The thesis aims to conduct a comprehensive review to analysis the role of AI in cyberattacks and its impact on red teaming practices.

Chapter 2 of the thesis analyses the submitted article, summarizing its methodology and outcomes. The article features a scoping review aimed to identify the AI methods employ in red teaming and the nature of their targeted attacks. Chapter 3 presents an extended literature review, which employs narrative review and snowball sampling methods to achieve its objectives. The review focuses on the applications of large language models (LLMs) used in red teaming attacks. It explores the role of LLMs and other advanced AI methods in the field of cyberattacks, with an emphasis on recent studies and their targets.

AI is driving transformative changes across the domain of red teaming and advanced AI such as LLMs offer both opportunities and risks. The rise of automated cyberattacks has introduced a new level of sophistication, making these attacks increasingly difficult to detect. Cybercriminals are leveraging accessible AI tools to execute automated and highly realistic attacks, often requiring minimal human intervention. These LLM based applications not only enable attackers to optimize their strategies but also present serious risks due to vulnerabilities within AI systems, potentially resulting in severe consequences. For instance, simulations of AI driven attacks have shown high success rates, highlighting the potential of these tools to enhance the methods of cyberattacks. Tools like Auto-GPT were discussed regarding their abilities for misuse if introduced to the public in the future. Research on AI in cyberattacks is needed to address the threats posed by the use of AI applications in red teaming.

Keywords

AI Red teaming, Artificial intelligence, Automated Red Teaming, LLMs, Generative AI, automated cyber attacks

Miscellaneous (Confidential information)

-

Contents

Figures

Tables

Acronyms

A3C	Advantage Actor-Critic
AI	Artificial Intelligence
ANN	Artificial Neural Network
CNN	convolutional neural network
DAE	Deep AutoEncoder
DDos	Distributed Denial-of-Service
DDQN	Double Deep Q-Network
DGA	Domain Generation Algorithms
DNN	Deep Neural Network
GA	Genetic Algorithm
GAN	Generative Adversarial Networks
GBRT	Gradient Boosting Regression Trees
GLRT	Generalized Likelihood Ratio Test
GPTs	Generative Pre-trained Transformers
GWO	Gray Wolf Optimization
KNN	k-Nearest Neighbors
LFA	Lagrangian Firefly Algorithm
LLMs	Large Language Models
LSTM	Long Short-Term Memory
LS-SVM	Least Squares Support Vector Machine
MLP	Multi-Layer Perceptron
NDAE	Nonsymmetric Deep AutoEncoder
NLP	Natural Language Processing
NSA	National Security Agency
OVA	One-Versus-All
PSO	Particle Swarm Optimization
RBM	Restricted Boltzmann Machine
RF	Random Forest
RL	Reinforcement learning
RLSC	Regularized Least-Squares Classification
RNN	Recurrent Neural Network

RWN	Random Weight Network
SDN	Software Defined Networking
SVC	Support Vector Classification
SVD	Singular Value Decomposition
SVM	Support Vector Machine
TOD	TensorFlow Object Detection
WorldCIST	World Conference on Information Systems and Technologies

1 Introduction

In the field of cybersecurity, more and more companies, individuals and governments are concerned about the harm that can caused by cyberattacks. As a result, they are putting a larger emphasis on their defence strategies to protect their infrastructure. On the other hand, attackers have been adopting new method to leverage cyberattacks by using AI based applications that lead cyberattacks with a high success rate and fast speeds.

The thesis discusses submitted article by examining the role of AI methods in leveraging cyberattacks and the typical targets. Appendix 1 includes the submitted article along with statements supporting its use as part of the thesis and as a published work. The extended section of the thesis focuses on the role of advanced AI in red teaming attacks, exploring the applications of LLM based systems. The main goal of the thesis is to extend the research to investigate the impact of LLM based applications such as GPT in cybercrimes. These tools have been used for a while now by everyone, but their threats are still undefined. Recent studies start to investigate these threats by examining how AI methods are used in red teaming, what the attacker aims to find, and the motivation behind these attacks. LLM based applications like ChatGPT-3 and ChatGPT-4 have demonstrated their ability to generate harmful behavior intentionally or unintentionally. These advanced LLM tools have the ability to collect large amounts of information about companies from their publicly shared data. Attackers could use this information for phishing attacks and sending malicious emails.

The review included recent studies that introduce LLM applications in their research as tools used in cyberattacks, but many of these LLM based applications have different targets and aim to conduct automated attacks. The studies simulate real cases where LLMs played a crucial role in red teaming attacks. The thesis investigates these attacks through a comprehensive literature review to examine AI methods used in these attacks, how the methods were employed, and whether the attacks were automated. The outcome of the thesis will define the misuse of AI tools in red teaming, identify their typical targets, and investigate the impact of these attacks.

1.1 Research Motivation

In recent years, the impact of AI on cybersecurity has grown significantly, particularly after 2022, when language models were introduced to the public, raising concerns about their ability to generate harmful behavior. The research examines various AI methods, focusing on how LLMs and other advanced AI methods are regularly used to execute attacks. The research aims to raise awareness about the role of AI in automated red teaming, the typical targets of these cyberattacks, and to introduce a better understanding of the impact of these attacks.

1.2 Research Objectives

The objective of the research is to examine the role of AI in red teaming through a comprehensive review, focusing on identifying the AI methods used in red teaming cyberattacks. The outcomes of the thesis will explore various AI techniques, tools, and frameworks used in red teaming and indicate the common methods introduced in these attacks, focusing on LLM based applications and exploring whether the attack was partially or fully automated. Furthermore, the research evaluates the goal of these AI technologies in enhancing the capabilities of red teaming and examines future trends in automated red teaming technologies, taking into account the advanced AI techniques that are changing the cyber threat landscape and their impact on cybersecurity measures.

1.3 Research Questions

Thesis questions are designed to address the uncover meaningful insights into the red teaming attacks and role of AI in an offensive context.

1. What is the role of LLM in red teaming?
2. What are the typical targets?
3. What risks do LLMs pose in cyberattacks?

1.4 Reliability and Ethics

The research followed the reliability and ethical framework by addressing relevant ethical issues and adhering to good scientific practice, using a methodology of data collection, analysis, and reporting. The research did not involve the processing of confidential or personal data.

The reliability of the thesis depends on the accuracy and integrity of the research, and the avoidance of plagiarism. According to Jamk (2024), plagiarism is a common form of misconduct associated with theses, involving the use of text from other individuals or artificial intelligence without proper permission or appropriate citations. The thesis applied good research practices with responsible inquiry and aimed to generate development competence within the results found by the thesis.

The thesis is supported by several factors that produce reliable research. A systematic methodology was employed to collect data using the scoping review method (Munn et al., 2018). The extended literature review followed a narrative method and snowball sampling. The scoping review followed a structured approach, analyzing articles from 2015 - 2023. A total of 471 articles were screened, but only 11 articles were included in the result of the article. The extended literature review examined 19 articles, focusing on advanced AI techniques such as LLMs in red teaming. The articles were aligned with the research criteria and provided answers to the research questions.

The databases for the sources used in the research were widely acknowledged in academic research such as Google Scholar and Janet Finna academic database, for a comprehensive review to cover the research objectives. The thesis also included figures and tables to enhance reader's understanding of the outcomes.

The research objectively adhered to its findings and followed Jamk's ethical principles and data protection guidelines (Jamk, 2024). The outcome of the thesis provides valuable results for the cybersecurity community and professionals.

1.5 Use of Artificial Intelligence

During the thesis process, ChatGpt-4 was used as an AI tool to help to produce good academic research and adhering to reliable resources. At the beginning, ChatGPT-4 was used for brainstorming ideas about the topic and asked to suggest resources related to the thesis topic. The results were unsatisfied because the references came from various databases which can be unclear or not adhering to the good scientific research principles. Later on, as the thesis progressed, ChatGPT-4 was used to check spelling and translation.

During the progress of the thesis, ChatGPT used to examine experiment to introduce harmful behavior so it can be added as experiment to define the threats associated with LLM based applications, but the result was unclear and did not add to results section.

Overall, the AI tool helped to better understand the purpose of the thesis and guidance in academic writing with consideration, following the Jamk's instruction on the use AI tools in writing thesis (Jamk, 2024).

2 Article summary: Artificial Intelligence in Red Teaming

This article was published by the authors Mays Al-Azzawi, Dung Doan, Tuomo Sipola, Jari Hautamäki and Tero Kokkonen in 2024. The research examined the impact of AI in red teaming exercises. The research used a scoping review methodology (Munn et al., 2018) through screening 471 articles to identify the impact of AI in cybersecurity landscape. Only 11 articles were included to align with the topic of the research to identify the AI methods used in red teaming exercises and the typical targets within these attacks (Al-Azzawi et al., 2024).

The article published in 2024 by Spring Nature, and it was included in the book series by Álvaro Rocha et al. The book is composed of several selected articles submitted to the World Conference on Information Systems and Technologies 2024 (WorldCIST 2024), as part of the Lecture Notes in Networks and Systems series (Álvaro Rocha et al.).

The article was submitted on November 10, 2023, and accepted on December 24, 2023. The registration was completed on January 5, 2024. Tero Kokkonen, one of the co-authors, presented the

findings at the WorldCIST 2024 conference, at Lodz University of Technology, Lodz, Poland. The research highlighted the potential risks and methodologies associated with AI in cybersecurity.

Tuomo Sipola, the corresponding author, played a crucial role in organizing the study and its submission. The work aligns with Jamk's focus on innovative technologies and reflects the collaborative efforts of the institution.

2.1 Purpose and Objective of the submitted Article

The article applied a scoping review method (Munn et al., 2018) to explore the role of AI in red teaming by examine the AI methods used in cyberattack activities and the targets for such of these attacks. The article highlighted the increasing threats generated by AI and assesses their risks.

The scoping review by Al-Azzawi et al. (2024) used a comprehensive literature review of existing articles between 2015 - 2023 provide a release of various types of AI techniques used in red teaming, and the target areas for such like attacks driven by AI. The research serves the cybersecurity community and professionals for better understanding and awareness of this threat.

2.2 The Research Methodology

The article applied the scoping review method (Munn et al., 2018) using Google Scholar and the Janet Finna academic database to search for related articles. The identification phase reviewed 471 articles related to the research objective through various keywords that aligned directly to define the scope of the topic. The research approach involved systematic mapping using an Excel sheet to sort information, identify the scope of the topic, and screen articles from the years 2015-2023. These articles were written in English. The first phase of the research involved identifying related articles by scanning their abstracts, introductions (in some chapters), and titles. The second phase was screening and excluding 460 articles that did not fit the article’s objectives. The final phase of the scoping review included only 11 articles, which were analyzed by reading them and determining whether they described the attack method, the target, how the attack was conducted, and the cyberattack methodology. The objective of the scoping review was to present the threat of AI. The submitted article met the criteria through its findings.

The scoping review applied a Prisma flow chart (McGowan et al., 2020) to their methodology to provide a clear understanding of the selection process the research. The Prisma flow presents the hierarchical structure of decomposition cycle undertake through the research phases. The chart divides the research to three phases which are identifying literature through different databases, screening these articles to define if they align with the research goal, and the final phase, included only the relevant articles to determine the result of the research.

2.3 Scoping Review Analysis

The research identified the role of AI in cyberattacks and their possible risk by responding to the research questions that defined the AI methods used in red teaming and the targets for such attacks. The identification phase revealed that articles from 2015 to 2018 which dealing with the topic of AI as attack tool, but it did not mention their method neither the targets. However, further studies presented a variety of attack methods, in Table 1 presents lists categories of AI methods mentions in the included articles which are 11 articles that relevant to the topic. Also, Table 1 appears in the thesis for clarification the results.

Table 1 List of methods discussed in the scoping review by Al-Azzawi et al. (2024)

Category	Methods	Description
Classification Methods	decision tree, convolutional neural network (CNN), recurrent neural network (RNN), long short-term memory (LSTM), support vector machine (SVM), support vector classification (SVC), deep neural network (DNN), least squares support vector machine (LS-SVM),	Methods for identifying categories or patterns (e.g., decision-making, image/text analysis).

	natural language processing (NLP), one-versus-all (OVA), double deep Q-network (DDQN), advantage actor-critic (A3C) regularized least-squares classification (RLSC), domain generation algorithms (DGA).	
Regression methods	Generative adversarial network (GAN), random forest (RF), multilayer perceptron (MLP), gradient boosting regression trees (GBRT), artificial neural network (ANN), logistic regression, generalized likelihood ratio test (GLRT)	Predicting continuous outputs (e.g., forecasting, trend analysis).
Clustering strategies	k-means clustering, restricted Boltzmann machine (RBM), particle swarm optimization (PSO), genetic algorithm (GA), deep autoencoder (DAE), Lagrangian firefly algorithm (LFA).	Identify patterns within data, which can be exploited for various purposes.
Other methods	nonsymmetric deep autoencoder (NDAE), cycle-GAN, combining TensorFlow object detection and a speech segmentation method with convolutional neural network (TOD+CNN), k-nearest neighbors (KNN), reinforcement learning (RL), gray wolf optimization (GWO), random weight network (RWN), ML-based approach named MLAPT, software defined networking (SDN), and singular value decomposition (SVD).	Advanced AI methods for specialized tasks (e.g., optimization, reinforcement learning).

The review of 11 related article introduced that LSTM was the most often method mentioned between these attacks, followed by Generative adversarial network (GAN) and support vector machine (SVM) in four literatures, as well as convolutional neural network (CNN), recurrent neural network (RNN), k-nearest neighbors (KNN), multilayer perceptron (MLP) and deep neural network (DNN) in three studies. Other methods were stated once or twice. Figure 1 was added to the thesis to provide a clear view of the frequent use of these methods within the 11 articles.

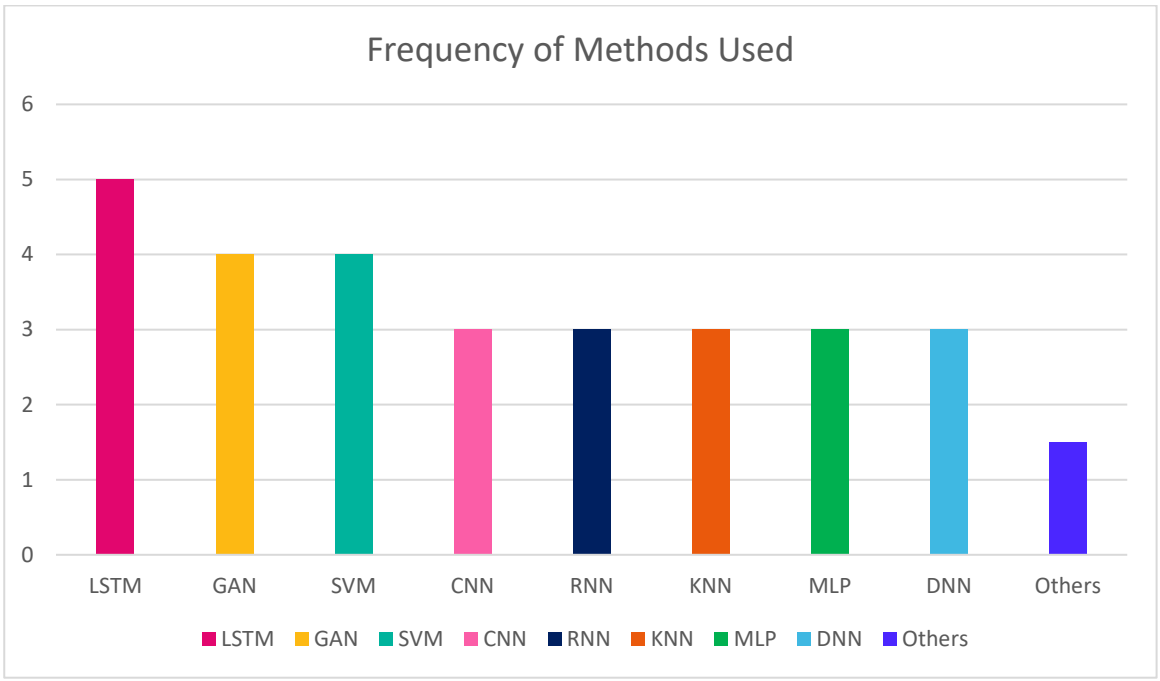

Figure 1 The frequency of the methods discussed in the reviewed articles

2.4 Discussion

The result identified various AI methods employed in red teaming. Among these methods long short-term memory (LSTM) was most frequently used, and the common targets were defined to expose sensitive data, passwords, social media profiles, and URLs.

The common cyberattack targets identified by Al-Azzawi et al. (2024) presents that the use of AI methods to collect data was the most frequently exploited by hackers, targeting areas such as personal data, sensitive data and health data, also, URLs, social media profiles, passwords and system information were also typical targets for such attacks. In Table 2 presents in the thesis to summarize the outcome of the targets that the attacker aims at.

Table 2 Summarize the frequency of AI attack targets and the attackers' purposes.

Category	Details	Frequency
General Data	Collect Health data, personal data, sensitive data (e.g., financial and government data).	Targeted in 4 articles.
URLs	Generate malicious URLs.	Targeted in 3 articles.
Social media profiles	Create fake profiles	Targeted in 2 articles.
Passwords	Testing user account passwords to gain access to the login.	Targeted in 2 articles.
Details of the systems	Fetch a sensitive information or files from the system	Targeted in 1 article.

The scoping review provides valuable insights into the impact of AI techniques on cyberattacks. The findings emphasize the methods employed and the target audiences. The risks associated with AI in red teaming are substantial and continue to expand as new techniques are developed to enhance the success rate of future cyberattacks."

3 Extended Literature Review

In this chapter, the research is extended from the original article to investigate the role of advanced AI methods in red teaming, focusing on recent studies that explore LLM based applications in executing these attacks. The extended comprehensive review found that LLM applications has important role in cyberattacks era. These applications used to plan attack strategies, but also

other studies investigated that LLM base applications like GPT has ability to leverage full automated attack. The targets of such attacks were to collect sensitive data from the target, create malicious URL, generate fake credential information, and applications like ChatGPT plays a critical role with social engineering attacks to generate harmful behavior.

3.1 Methodology

In order to identify relevant articles related to the topic, a comprehensive literature review was conducted. Phase 1 involved a narrative review methodology (Sukhera, 2022) on Google Scholar and the Janet Finna academic database, using various keywords to identify articles related to the topic. The outcome of Phase 1 included only 5 articles. In Phase 2, the research employed snowball sampling (Naderifar et al., 2017) to explore citations from the initially identified sources, resulting in the inclusion of 14 articles related to the objectives of the thesis. The search followed a flexible and accurate approach to analyze 19 articles written in English.

After gathering the most relevant information from various sources using different methods, the result of the extended review focused on addressing the research questions and identifying the common LLM techniques used in red teaming within frameworks and systems to investigate the targets of these attacks. The gathered results emphasize the role of different AI techniques in automated cybercrimes, the use of LLMs in red teaming, and how LLMs are employed to simulate automated attacks.

3.2 Results

After the identification and screening phases of the extended part was conducted through comprehensive review exploring the role of LLM in red teaming. Yuen (2015) presented a concept of automated planning and how to employ this principle in automated cyber red teaming. The automated plan is a branch of AI that generates step by step plan to solve problem, and the result was to draft attack plan to identify vulnerabilities in network systems using tools and frameworks for automation which these kind of experiment helps in improves the defences strategies.

Waizel (2024) stated many artificial intelligence methods used in automated cyberattacks by automate the process of testing stolen credentials on websites using credential stuffing method. Pal et

al. (2019) proposed the credential stuffing method using natural networks such of these applications using recurrent neural networks (RNNs) and deep generative adversarial networks (GAN) to generate passwords (Pal et al., 2019). Waizel (2024) introduced Botnets, AI powered bots used as attack method. Botnets attacks could lead a destituted denial of service (DDos) attacks, spreading malware or conducting large-scale spam campaigns (Waizel, 2024). For depth understanding to A based Botnets, Yang and Menczer (2023) presented how LLM based bots like ChatGPT that can automatically create fake profiles and used to spread malicious, other examples of these bots like fox8 botnet used to create faked profiles on X platform (Twitter, known formerly) (Yang & Menczer, 2023).

In 2022 language models were publicly revealed and since that a new era of cybercrimes has been found. Many automated red teaming exercises have leveraged LLMs in their strategies (Mazeika et al., 2024). Perez and Riberio (2022) proposed attacks methods that demonstrate how LLMs Like GPT-3 can misused by attackers to generate malicious. Hazell (2023) discussed how LLM based applications like GPT can be used for automated social attacks by creating spear phasing attacks to generate malicious emails. Brundage et al. (2018) stated that according ZeroFox's researchers, they investigated the adoption of a fully automated spear phishing system on social media platforms by creating tweets that highly rate of clicked on malicious links (Brundage et al., 2018).

Hendrycks et al. (2023) addressed various categories of AI-driven threats and how AI can lead automated attack. The authors discussed a scenario based on information released from the national security agency (NSA) in 2014 that they are developing a system called MonsterMind, which automated to identify and prevent the cyberattacks for Unites States infrastructure, but the concern was that, in future it could be used to lead cyberattacks without human interaction. The authors discussed the threat behind the policies of employing automation attacks and how a mistake could lead to genuine conflict without human awareness (Hendrycks et al., 2023).

Attack scenario demonstrated by Pasquini et al. (2024) using LLM agent as attacker in an experiment. The LLM agent execute automated cyberattacks by planning and execution. The LLM agent improves its strategy based on feedback collected from the target. The LLM agent define to be combined with a framework that allows it to autonomously engage with its environment (Yao et al., 2022). The experiment produced by Pasquini et al. (2024) that LLM agent used as backend to

execute attack using OpenAI and Anthropic such as ChatGPT-4 and Claud3.5-Sonnet to execute these attacks.

Studies by Xu et al. (2024) introduced the AutoAttacker framework to leverage fully automated attack to gain access to the target system, as shown in Figure 2, the framework used LLMs to automate the attack without the need to human, and the framework generates a simulation of complex social engineering and phishing strategies. AuttoAttacker employed reinforcement learning (RL) to improve attack strategies, depending on collecting feedback from the target responses (Tsingenopoulos et al., 2019; Xu et al., 2024). The performance of the AutoAttacker shows high effective attack especially when using GPT-4 (Xu et al., 2024).

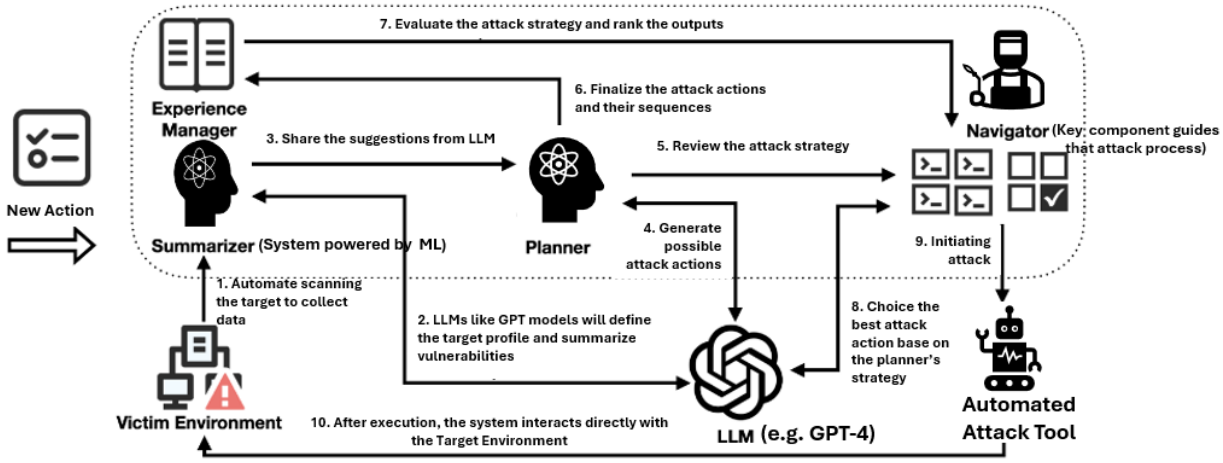

Figure 2 Adapted from Xu et al. (2024) to explain the workflow of AutoAttacker and the role of LLM in generating and excute automated attack.

The LLM based application may not be directly related to automated red teaming, but it could play a vital part as Dhamani (2024) introduces a few scenarios where AI is misused, scenarios engaged with advanced LLMs used to introduce phishing attacks using ChatGPT by spread phishing email. ChatGPT was able to create emails based on the company published data to draft emails looks good as the originals.

Below in Table 3, presents the role of AI and LLMs techniques used in cyberattacks and automated attacks. The table presents that LLM based applications are the most common between other techniques. Key findings highlight how AI systems, such as MonsterMind, and LLM applications like GPT-3 and GPT-4, are leveraged to plan, execute, and enhance cyberattacks. Techniques include credential stuffing using neural networks, social engineering, phishing, and automated attacks through frameworks like AutoAttacker. Additionally, LLM powered bots facilitate malicious activities by generating realistic phishing emails, fake social media profiles, and malicious links. These results demonstrate the increasing sophistication and automation of cyberattacks driven by AI technologies.

Table 3 AI and LLM Techniques in Automated Red Teaming and Cyberattacks

AI in automated cyberattacks	**LLMs in red teaming**	**LLMs in automated cyberattacks.**
MonsterMind: AI system revealed by the NSA in 2014 that could perform fully automated attacks. (Hendrycks et al., 2023)	Botnets: LLM based bots like ChatGPT can be used to generate fake profiles and assist in spreading malicious content (Yang & Menczer, 2023; Waizel, 2024).	Automated Social Media Attacks: LLM based bots can create tweets and posts designed to trick users into clicking on malicious links (Brundage et al., 2018).
Credential Stuffing: Automated testing of stolen credentials on websites, using RNNs and GANs to generate passwords (Pal et al., 2019; Waizel, 2024).	Spear Phishing: LLMs such as GPT can generate convincing spear-phishing emails by mimicking legitimate communication styles (Hazell, 2023).	Agent Attacks: LLMs like GPT-4 and Claude 3.5-Sonnet are used within agent attacks to autonomously plan and execute strategies, adapting based on feedback from the target (Pasquini et al., 2024).
	Social Engineering and Phishing: LLMs craft persuasive	AutoAttacker Framework: LLMs like GPT-4 are used to simulate

	phishing emails and messages by analyzing context and producing realistic content (Dhamani, 2024).	sophisticated social engineering and phishing attacks, improving strategies through reinforcement learning (Xu et al., 2024; Tsingenopoulos et al., 2019). Automated Red Teaming: LLMs like GPT-3 and GPT-4 simulate attacks, generate strategies, and improve their execution autonomously (Perez & Ribeiro, 2022; Mazeika et al., 2024).

The common targets of these cyberattacks have been identified as follows:

- Login and Credential-Based Attacks: Automated credential stuffing using stolen credentials, AI-driven password generation, and botnets for system access (Pal et al., 2019; Waizel, 2024).
- URLs: Malicious links embedded in phishing emails or social media posts to deceive users (Brundage et al., 2018).
- Social Media Profiles: Fake profiles created by bots for malicious activities, such as spreading misinformation (Yang & Menczer, 2023; Waizel, 2024).
- Passwords: AI-driven tools like GANs and RNNs used to generate and exploit user account passwords (Pal et al., 2019).
- Detail of Systems: Exploitation of system architecture or operational information for automated attacks (Xu et al., 2024; Hendrycks et al., 2023).
- Emails: Phishing emails mimicking legitimate corporate communication using LLMs (Hazell, 2023; Dhamani, 2024).

3.3 Discussion

Artificial intelligence applications have transformed our perspective for everything and the impact of AI in all domains. The cyberattacks in response to these changes are increasing. Attackers taking advantage of these tools that are available to everyone. As a result, the game is changing by using these techniques should be considered especially in cybersecurity landscape. The ongoing development of generative pre-trained transformers (GPTs) applications successfully leveraging more realistic attacks that harder to detect.

The risk of conducting these automated cyberattacks and enabling advanced LLM based applications to make decisions lies in their potential, this risk could cause significant harm. The automated AI methods can lead to severe consequences, as these applications are vulnerable to faults within their results.

Within all, these challenges and threats in cybersecurity era, AI based applications is taking the role in driving cyberattacks in fast speed and may not need to human interaction. These challenges clearly identified the optimization of the future of cyberattacks. Attackers obtained LLM in their strategies to improve the strategies plan of the attacks and employed these methods in automated cyberattacks, the result produce highly success rate in achieving the attacks. Recent studies have simulated experiments to identify the threats associated within LLM based applications, using advanced LLM in frameworks like GPT-3 or GPT-4 to achieve these attacks. These applications in continuous developing to enhance their results and the future promises in Auto-GPT, these AI methods have no limitation to their abilities within cybercriminals, The future research should evaluate these threats and examine the result within LLM applications. The researchers are still observing these tools and examining their vulnerabilities.

4 Conclusion

This thesis explored the impact of artificial intelligence (AI) in red teaming exercises, focusing on identifying the methods used and the typical targets of these attacks. The research was guided by these questions regarding to explore the LLM techniques used in red teaming, the attacker’s aims, and the associated risks. By conducting a scoping review and narrative review to extend the research, the study revealed significant insights into the capabilities of AI methods in executing

cyberattacks. LLM based application show their ability in cyberattacks, such as phishing, credential stuffing, and generating harmful social media content. Key findings also highlighted the frequent use of LLM tools like ChatGPT, and other AI methods such as frameworks in automated red teaming.

The reliability of the study was supported by a systematic methodology, including collect data and adherence to ethical principles. There was relatively small sample size of relevant articles in the comprehensive review of the findings. Additionally, the complexity in distinguishing fully automated attacks and human led the attacks, introduces challenges in fully evaluating the risks of LLM based application. These factors underline the need for further research.

Researcher should investigate deeper into understanding the impact of emerging AI techniques, such as Auto-GPT, and their potential misuse in cyberattacks. It should also explore effective defense mechanisms, policy implications, and the development of tools to mitigate the risks caused by AI driven cyber threats. By addressing these issues, future studies can contribute to enhancing the cybersecurity scene.

References

Al-Azzawi, M., Doan, D., Sipola, T., Hautamäki, J., & Kokkonen, T. (2024, March). Artificial Intelligence Cyberattacks in Red Teaming: A Scoping Review. In *World Conference on Information Systems and Technologies* (pp. 129-138). Cham: Springer Nature Switzerland.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., ... & Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*.

Dhamani, N. (2024). *Introduction to Generative AI*. Manning Publications Co. LLC.

Hazell, J. (2023). Spear phishing with large language models. *arXiv preprint arXiv:2305.06972*.

Hendrycks, D., Mazeika, M., & Woodside, T. (2023). An overview of catastrophic ai risks. *arXiv preprint arXiv:2306.12001*.

Jamk. (2024). *Use of artificial intelligence in learning assignments and thesis*. JAMK University of Applied Sciences. https://help.jamk.fi/raportointiohje/en/3-the-writing-process/use-of-artificial-intelligence-in-learning-assignments-and-thesis/

Lin, L., Mu, H., Zhai, Z., Wang, M., Wang, Y., Wang, R., ... & Li, H. (2024). Against The Achilles' Heel: A Survey on Red Teaming for Generative Models. *arXiv preprint arXiv:2404.00629*.

Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., Mu, N., ... & Hendrycks, D. (2024). Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*.

McGowan, J., Straus, S., Moher, D., Langlois, EV, O'Brien, KK, Horsley, T., ... & Tricco, AC (2020). Reporting scoping reviews—PRISMA ScR extension. *Journal of clinical epidemiology*, *123*, 177-179.

Munn, Z., Peters, MD, Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC medical research methodology*, *18*, 1-7.

Naderifar, M., Goli, H., & Ghaljaie, F. (2017). Snowball sampling: A purposeful method of sampling in qualitative research. *Strides in development of medical education*, *14*(3).

Pal, B., Daniel, T., Chatterjee, R., & Ristenpart, T. (2019, May). Beyond credential stuffing: Password similarity models using neural networks. In *2019 IEEE Symposium on Security and Privacy (SP)* (pp. 417-434). IEEE.

Pasquini, D., Kornaropoulos, E. M., & Ateniese, G. (2024). Hacking Back the AI-Hacker: Prompt Injection as a Defense Against LLM-driven Cyberattacks. *arXiv preprint arXiv:2410.20911*.

Perez, F., & Ribeiro, I. (2022). Ignore previous prompt: Attack techniques for language models. *arXiv preprint arXiv:2211.09527*.

Rocha, Á., Adeli, H., Dzemyda, G., Moreira, F., & Poniszewska-Marańda, A. Good Practices and New Perspectives in Information Systems and Technologies

Sukhera J. (2022). Narrative Reviews in Medical Education: Key Steps for Researchers. *Journal of graduate medical education*, *14*(4), 418–419. https://doi.org/10.4300/JGME-D-22-00481.1

Tsingenopoulos, I., Preuveneers, D., & Joosen, W. (2019, June). AutoAttacker: A reinforcement learning approach for black-box adversarial attacks. In *2019 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)* (pp. 229-237). IEEE.

Waizel, G. (2024, July). Bridging the AI divide: The evolving arms race between AI-driven cyber-attacks and AI-powered cybersecurity defenses. In *International Conference on Machine Intelligence & Security for Smart Cities (TRUST) Proceedings* (Vol. 1, pp. 141-156).

Xu, J., Stokes, J. W., McDonald, G., Bai, X., Marshall, D., Wang, S., ... & Li, Z. (2024). Autoattacker: A large language model guided system to implement automatic cyber-attacks. *arXiv preprint arXiv:2403.01038*.

Yang, K. C., & Menczer, F. (2023). Anatomy of an AI-powered malicious social botnet. *arXiv preprint arXiv:2307.16336*.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2022). React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Yuen, J. (2015). Automated cyber red teaming. *Cyber and Electronic Warfare Division, Defence Science and Technology Organisation, Edinburgh South Australia, Australia, Tech. Rep*.

Appendices

Appendix 1. Submitted article as pdf and the permission to conduct this thesis

The thesis includes material reproduced with permission from Spring Nature:

Al-Azzawi, M., Doan, D., Sipola, T., Hautamäki, J., & Kokkonen, T. (2024). Artificial intelligence cyberattacks in red teaming: A scoping review. In Á. Rocha, H. Adeli, G. Dzemyda, F. Moreira, & A. Poniszewska-Marańda (Eds.), *Good practices and new perspectives in information systems and technologies*: WorldCIST 2024 (Lecture Notes in Networks and Systems, Vol. *985*, pp. 129–138). Springer, Cham. https://doi.org/10.1007/978-3-031-60215-3_13

First published in Good Practices and New Perspectives in Information Systems and Technologies, 129–138, 2024, by Springer Nature.

Artificial Intelligence Cyberattacks in Red Teaming: A Scoping Review

Mays Al-Azzawi[1], Dung Doan[2], Tuomo Sipola[3(✉)], Jari Hautamäki[3], and Tero Kokkonen[3]

Institute of Information Technology,
Jamk University of Applied Sciences,
Jyväskylä, Finland
[1]ab0168@student.jamk.fi
[2]aa7785@student.jamk.fi
[3]{tuomo.sipola, jari.hautamaki, tero.kokkonen}@jamk.fi

Abstract. Advances in artificial intelligence are creating possibilities to use these methods in red team activities, such as cyberattacks. These AI attacks can automate the process of penetrating a target or collecting sensitive data while accelerating the pace of carrying out the attacks. This survey explores how AI is employed in cybersecurity attacks and what kind of targets are typical. We used scoping review methodology to sift through articles to find out AI methods, targets, and models that red teams can use to emulate cybercrime. Out of the 470 records screened, 11 were included in the review. Multiple cyberattack methods can be found to exploit sensitive data, systems, social media user profiles, passwords, and URLs. The use of AI in cybercrime to build versatile attack models poses a growing threat. Additionally, cybersecurity can use AI-based techniques to offer better protection tools to deal with those problems.

Keywords: artificial intelligence, red team, red teaming, cyberattack, cybersecurity

1 Introduction

The landscape of cybersecurity has undergone an enormous change in the last few years. One phenomenon that stands out is the possibility of artificial intelligence simulating human behavior. The behavior of artificial intelligence in cybersecurity can lead to dangerous situations in terms of security. Using AI as a method for attacks has developed in tandem with the development of attack methodologies and AI capabilities. Only a few cases are reported, and simulating human acts has become more feasible in the last few years.

The term red teaming originates from the military domain as a way to role-play adversaries or assess vulnerabilities [12]. The term red team also originates from widely used military symbols such as APP-6 by NATO or MIL-STD-2525 by U.S. Department of Defense, where the hostile (and suspect) identity is indicated with a red color [15,1]. In the context of cybersecurity, U.S. National Institute

The original article appeared as: Mays Al-Azzawi, Dung Doan, Tuomo Sipola, Jari Hautamäki and Tero Kokkonen. "Artificial Intelligence Cyberattacks in Red Teaming: A Scoping Review." In: *Good Practices and New Perspectives in Information Systems and Technologies. WorldCIST 2024.* Ed. by Álvaro Rocha, Hojjat Adeli, Gintautas Dzemyda, Fernando Moreira, and Aneta Poniszewska-Marańda. Vol. 1. Lecture Notes in Networks and Systems 985. Cham, Switzerland: Springer, 2024, pp. 129–138. https://doi.org/10.1007/978-3-031-60215-3_13

of Standards and Technology (NIST) defines a red team as follows: *"A group of people authorized and organized to emulate a potential adversary's attack."* [5] The red teams improve enterprise security by demonstrating the impacts of successful attacks [5]. In the context of cybersecurity, the term red team is used in cybersecurity exercises and in security testing. In cybersecurity exercises, red teams (RT) simulate the threat actors of the exercise scenario by executing cyberattacks against blue teams (BT), which are defending their assets [11,3,24,9,19]. In security testing, the red team is the group of security testers.

AI red teaming can be understood as an activity from two different perspectives. Several large technology companies use red teaming to expose weaknesses and vulnerabilities in their systems [18,27]. Another aspect is the use of AI to carry out attacks, which can be targeted against technical systems. On the other hand, in social engineering-type attacks, AI is used as a stepping stone to advanced persistent threat (APT) attacks by searching for suitable victims that can be targeted by AI-generated ghost messages [6,17]. The advantage of AI specifically in such attacks is the ability to enable mass attacks using phishing techniques to open attack vectors to multiple targets instead of manual attacks. For example, AI-generated phishing messages in target language create persuasive attack vectors. AI-based solutions are built to make operations more effective. Automating the process of planning attacks for automated cybersecurity testing scenarios could save time and effort [26]. As new artificial intelligence technologies have become more prevalent, automation is easier to implement, although its impact on work and society should be studied [22].

In order to investigate the use of AI for cyberattacks for red teaming, we carried out a scoping review. To examine how AI can be used for cyberattacks, red team actions, and hacking, our research questions were the following:

- *RQ1:* What AI attack methods are there?
- *RQ2:* What are the targets of such attacks?

Next, this paper describes the used scoping review methodology in Section 2, including a figure of the review protocol. The results of the review are presented in Section 3 with two tables summarizing the main findings. Finally, a conclusion is provided in Section 4.

2 Methodology

We used the scoping review method [14] to search the academic Finna[1] library database and Google Scholar[2] in order to define the scope of our topic. The review considered the following keywords: ' defensive mission', 'AI-enabled cyber operations', 'AI-augmented cyber defenses', 'national defense postures', 'poisoning attacks', 'offensive cyber operations', 'Cyber activities', 'AI cyber operations', 'AI cyber defense', 'AI cyber attack', 'AI red teaming', 'AI-enabled cyber

[1] https://janet.finna.fi/
[2] https://scholar.google.com/

campaigns', and 'cyber attacks'. In the initial stage, we identified 471 articles (and some book chapters) by screening their titles and abstracts within the 2015–2023 timeframe, found at the time of the research in mid-2023. We included articles written in English with available abstracts. During the second phase of the research, a more involved analysis of these articles was conducted. This analysis included reading the articles closely and concentrating on the topic at hand to precisely determine their content and classify them as directly relevant to addressing the research questions (RQ1, RQ2). We used the following criteria to find answers:

1. Is there a description of an attack method?
2. What was the target of the attacks?
3. How was the attack conducted?
4. What was the cyber-attack methodology used?

The result of the second stage of the research yielded 11 articles related to the subject matter. In the third stage of the research, we composed summaries, which also involved addressing the aforementioned questions when applicable. This comprehensive analysis of the included studies enabled us to gather information on the utilization of AI in red teaming. The review process is detailed by the PRISMA flow chart [13] in Figure 1.

3 Results

3.1 Attack methods

The literature review encompassed studies published from 2015 to 2023 in which we identified various cyberattack methods. The following techniques were documented in those studies. *Classification methods:* decision tree, convolutional neural network (CNN), recurrent neural network (RNN), long short-term memory (LSTM), support vector machine (SVM), support vector classification (SVC), deep neural network (DNN), least squares support vector machine (LS-SVM), natural language processing (NLP), one-versus-all (OVA), double deep Q-network (DDQN), advantage actor-critic (A3C) regularized least-squares classification (RLSC), domain generation algorithms (DGA). *Regression methods:* generative adversarial network (GAN), random forest (RF), multilayer perceptron (MLP), gradient boosting regression trees (GBRT), artificial neural network (ANN), logistic regression, generalized likelihood ratio test (GLRT), *Clustering strategies:* k-means clustering, restricted Boltzmann machine (RBM), particle swarm optimization (PSO), genetic algorithm (GA), deep autoencoder (DAE), Lagrangian firefly algorithm (LFA), *Other specific methods:* nonsymmetric deep autoencoder (NDAE), cycle-GAN, combining TensorFlow object detection and a speech segmentation method with convolutional neural network (TOD+CNN), k-nearest neighbors (KNN), reinforcement learning (RL), gray wolf optimization (GWO), random weight network (RWN), ML-based approach named MLAPT, software-defined networking (SDN), and singular value decomposition (SVD).

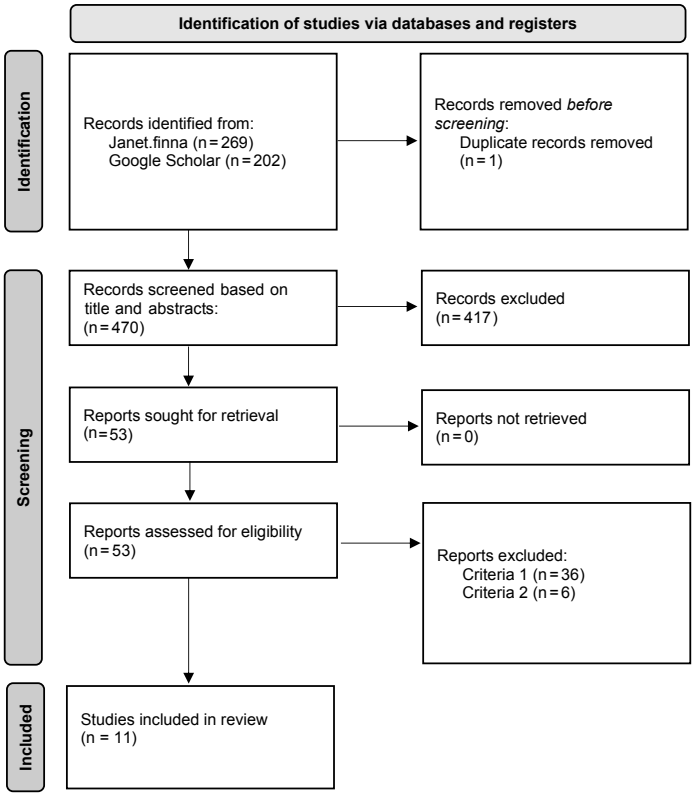

Fig. 1. Review protocol

Among these methods, LSTM was the most frequently used, appearing in 5 of the reviewed articles, while GANs and SVM were employed in 4 studies each. Additionally, CNN, RNN, KNN, MLP, and DNN were each featured in three of the reviewed articles. Other methods were referenced only once or twice. For a list of the attack methods in the reviewed articles, refer to Table 1.

3.2 Attack targets

Furthermore, we identified common targets that cyberattackers typically aim at (see Table 2 for tabulation of targets), including:

- General data, such as health data, personal data, and sensitive data, including financial and government data, were the most frequently targeted, appearing in 4 of the reviewed articles. [2] [20] [28] [25]
- URLs: Attackers also frequently targeted URLs, with 3 instances in the reviewed articles. [10] [28] [7]
- Social media user profiles: This category was the target in 2 of the sources. [10] [7]
- Passwords: Passwords were a target in 2 sources. [28] [7]
- Details of systems: Details of systems were targeted in one article. [25]

3.3 Summaries

The use of AI has been identified as a cyberattack method and recognized as a potential risk. However, Clinton only presents AI as a hacking method, and we did not find any other specific attack methods. [4]

Ward et al. have defined artificial intelligence as a new technology used by hackers and have mentioned "poison" attacks utilizing machine learning algorithms. They also discuss automated vehicles and the potential for high-risk attacks on vehicle systems. However, during their discussion of AI hacking methods, no specific attack methods were mentioned. [23]

From 2015 to 2018, the articles about AI-hacking did not mention any attack methods, and targets were mainly data and sensitive data.

Yamin et al. focused on raising awareness about the use of artificial intelligence as an attack method and assessed its impact on military operations. They employed GANs and Nash equilibrium to describe the attack methods. The targets of these attacks included traffic signs, medical image data, facial image data, digital recommendation systems, CT-scan data, speech and audio data, as well as network intrusion detection systems. The attacks were carried out using malicious AI algorithms designed to manipulate data to evade benign AI algorithm classifiers. The methodologies employed in these cyberattacks included DeepHack, DeepLocker, GyoiThon, EagleEye, Malware-GAN, UriDeep, Deep Exploit, and DeepGenerator. [25]

The article by Kaloudi et al. investigates AI's threat to SCPS. It explores how AI can be used as a malicious tool, emphasizing its potential to increase

Table 1. Methodologies found in the reviewed articles.

Author	Pistono and Yampolskiy	Brundage et al.	Kaloudi and Li	King et al.	Truong et al.	Zouave et al.	Yamin et al.	Wang et al.	Guembe et al.	$\sum$
Year	2016	2018	2020	2020	2020	2020	2021	2022	2022	
Reference	[16]	[2]	[8]	[10]	[20]	[28]	[25]	[21]	[7]	
Classification										
Dec. tree						x				1
CNN					x	x			x	3
RNN			x			x			x	3
LSTM			x		x	x		x	x	5
SVM					x	x		x	x	4
SVC						x			x	2
DNN			x					x	x	3
LS-SVM					x					1
NLP						x				1
OVA						x				1
DDQN								x		1
A3C								x		1
RLSC						x				1
DGA						x				1
Regression										
GANs						x	x	x	x	4
RF						x			x	2
MLP					x	x			x	3
GBRT						x			x	2
ANN					x					1
Log. reg.						x				1
GLRT					x					1
Clustering										
k-means			x							1
RBM					x					1
PSO					x					1
GA					x					1
DAE					x					1
LFA					x					1
Other										
NDAE					x					1
CYCLE-GAN									x	1
TOD+CNN									x	1
KNN					x	x			x	3
RL			x							1
GWO					x					1
RWN					x					1
MLAPT					x					1
SDN								x		1
SVD								x		1

Table 2. Attack targets found in the reviewed articles.

Author	Pistono and Yampolskiy	Brundage et al.	Kaloudi and Li	King et al.	Truong et al.	Zouave et al.	Yamin et al.	Wang et al.	Guembe et al.	$\sum$
Year	2016	2018	2020	2020	2020	2020	2021	2022	2022	
Reference	[16]	[2]	[8]	[10]	[20]	[28]	[25]	[21]	[7]	
Data, sensitive data		x			x	x	x			4
URLs				x		x			x	3
Social media user profiles				x					x	2
Password						x			x	2
Systems							x			1

attack speed and success rates. Attack methods discussed include k-means clustering, RNN, LSTM, RL, and DNN. Case studies involve k-means clustering for phishing messages, RNN for deceptive reviews, LSTM for phishing URLs, RL for autonomous learning attacks, and DNN for cyberattacks. The paper also examines cyberattack methodologies, including DeepLocker, repurpose attacks, DeepHack, Deep-Phish, review attacks, and SNAP_R. [8]

Guembe et al. address the growing concern of AI-powered cyberattacks and provide insights into how AI can be maliciously utilized in such attacks. They employ various attack methods, including CNN, GAN, RNN, LSTM, SVC, SVM, cycle-GAN, TOD+CNN, RF, MLP, GBRT, KNN, and DNN. The targets of these attacks encompass public social media profiles, passwords, and URLs. The attacks are executed through techniques such as password guessing/cracking (brute-force attacks), intelligent captcha manipulation, smart abnormal behavioral generation, AI model manipulation, and the generation of sophisticated fake reviews. The cyberattack methodologies employed by the authors include DeepLocker, DeepHack, PassGAN, and HashCat. [7]

Truong et al. provide an insightful overview of how artificial intelligence can be leveraged in cybersecurity, both for offensive and defensive purposes. They employ a diverse set of attack methods, including SVM, RBM, MLP, KNN, CNN, PSO, GA, DAE, ANN, LS-SVM, NDAE, GWO, RWN, LFA, MLAPT, LSTM, and GLRT. The targets of these attacks encompass user identities, financial credentials, and sensitive data from large corporations, security agencies, and government organizations. These attacks serve various purposes, including detecting or categorizing malware, identifying network intrusions, countering phishing and spam attacks, mitigating Advanced Persistent Threats (APTs), and identifying domains generated by domain generation algorithms (DGAs). [20]

Articles in 2020 showed different attack methods, such as GANs, CNN, RNN, LSTM, SVM, and SVC, aimed at attacking sensitive data, social media user profiles, passwords, and URLs.

The article by Zouave et al. explores the possibilities and applications of AI throughout various stages of a cyberattack. The authors employ a wide range of attack methods, including RNN, LSTM, NLP, GAN, KNN, logistic regression, SVC, decision tree, RF, gradient boosting regression tree, SVM, MLP, RLSC, OvA, CNN, and DGA. These attacks target URLs, individuals' personal data

in search of relationships, passwords, captchas, and domains. The attacks are executed by creating deceptive URLs to evade automated detection, generating conversations that include harmful links and attachments, attempting password guessing and brute forcing, stealing passwords, solving captchas, and generating numerous random fake domains. The authors utilize cyberattack methodologies such as the DeepPhish algorithm, PassGAN, Torch RNN, Deeptcha, AGDs, and DeepDGA. [28]

In the article by Wang et al. the exploration focuses on poisoning attacks in machine learning, particularly within the context of automated vehicles. The authors utilize various attack methodologies harnessing AI techniques. These include deep learning and deep neural networks (DNN), known for their outstanding performance in recognition tasks like image classification and computer vision. Additionally, other methods are discussed, such as Generative Adversarial Networks (GAN), LSTM, SDN, DDQN (Deep Double Q-Network), Advantage Actor-Critic (A3C), SVM (Support Vector Machine), and singular value decomposition (SVD). [21]

The article by Brundage et al. provides a summary of workshop findings and the authors' conclusions on forecasting, preventing, and mitigating the detrimental impacts of malicious AI use. The targets included sensitive information or financial assets of individuals, specific members of crowds, and historical patterns of code vulnerabilities. The attacks were executed through various methods, such as spear phishing attacks, imitation of human-like behavior, facial recognition, the generation of custom malicious websites/emails/links, visual impersonation of another person in video chats, and the use of drones or autonomous vehicles to deliver explosives and cause accidents. Furthermore, the attackers were engaged in discovering new vulnerabilities and developing code to exploit them. However, no specific methods for these activities were mentioned in the report. [2]

In their article, King et al. introduced the term "AI-Crime" (AIC) to address two key questions regarding the threats posed by AI in criminal activities and potential solutions to mitigate these threats. However, the article does not specify the methods employed in these AIC activities. The primary target of these activities is social media users, particularly through the use of phishing links. [10]

The research paper by Pistono and Yampolskiy focuses on publishing papers related to malicious exploits and discusses the use of software with malicious capabilities, including truly artificially intelligent systems such as artificially intelligent viruses. The paper also introduces the term "Hazardous Intelligent Software" (HIS) to describe the use of intelligence in a malicious context. It highlights that intelligent systems can potentially become malevolent in various ways. However, the paper does not mention specific AI attack methods. [16]

4 Conclusion

In today's rapidly evolving digital landscape, cybercriminals are continuously adapting and enhancing their attack strategies, with a particular focus on lever-

aging AI-driven techniques. Our results indicate that primary targets (RQ2) include personal data as well as sensitive information held by governments, organizations, and individuals, spanning URLs, passwords, and critical systems. Furthermore, the results show that to achieve their malicious goals (RQ1), cybercriminals can exploit a wide array of machine learning methods, falling into distinct categories: Classification, Regression, and Clustering.

These categories contain various technologies used for attacks. *Classification Techniques:* They include a multitude of machine learning algorithms such as decision trees, CNN, RNN, LSTM, SVM, SVC, DNN, LS-SVM, NLP, OVA, DDQN, A3C, RLSC, and DGA. These methods enable cybercriminals to classify and categorize data, often to identify vulnerabilities or potential targets. *Regression Methods:* In this category, we find techniques such as GANs, RF, MLP, GBRT, MLP, ANN, Logistic Regression, and GLRT. These approaches are employed to predict and estimate various variables, ranging from password guessing to system security breaches. *Clustering Strategies:* Cybercriminals also rely on clustering methods such as k-means clustering, RBM, PSO, GA, DAE, and LFA. Clustering helps them identify patterns within data, which can be exploited for nefarious purposes.

Cybercriminals employ sophisticated methodologies like the DeepPhish algorithm, PassGAN, Torch RNN, and Deeptcha. These tools aid them in tasks such as cracking passwords, phishing attacks, and infiltrating secure systems. As the threat landscape continues to evolve, it is imperative for the security research community, government agencies, and cybersecurity experts to remain vigilant and well-prepared against AI-based attacks. Red teaming using these AI-based attacks could reveal vulnerabilities to novel attacks. Effective countermeasures and proactive strategies must be developed to address the growing challenges posed by AI-driven cyberattacks.

Acknowledgements

This research was partially funded by the Resilience of Modern Value Chains in a Sustainable Energy System project, co-funded by the European Union and the Regional Council of Central Finland (grant number J10052). The authors would like to thank Ms. Tuula Kotikoski for proofreading the manuscript.

References

1. Department of defence interface standard, common warfighting symbology. Standard MIL-STD-2525C, United States of America, Department of Defence (2008)
2. Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., et al.: The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv preprint arXiv:1802.07228 (2018)
3. Brynielsson, J., Franke, U., Tariq, M.A., Varga, S.: Using Cyber Defense Exercises to Obtain Additional Data for Attacker Profiling. In: 2016 IEEE Conference on Intelligence and Security Informatics (ISI), pp. 37–42 (2016). DOI 10.1109/ISI.2016.7745440

4. Clinton, L. (ed.): Cybersecurity for business. Kogan Page, London, England (2022)
5. Computer Security Resource Center (CSRC) of National Institute of Standards and Technology (NIST): The glossary of terms and definitions extracted verbatim from nist's cybersecurity- and privacy-related publications. URL https://csrc.nist.gov/glossary/term/red_team. Accessed: 15 September 2023
6. Ghafir, I., Prenosil, V.: Advanced persistent threat and spear phishing emails. In: M. Hrubý (ed.) Proceedings of the International Conference Distance Learning, Simulation and Communication 'DLSC 2015', pp. 34–41. University of Defence, Brno, Czech Republic (2015)
7. Guembe, B., Azeta, A., Misra, S., Osamor, V.C., Fernandez-Sanz, L., Pospelova, V.: The emerging threat of ai-driven cyber attacks: A review. Applied Artificial Intelligence **36**(1), 2037,254 (2022)
8. Kaloudi, N., Li, J.: The AI-based cyber threat landscape: A survey. ACM Computing Surveys (CSUR) **53**(1), 1–34 (2020)
9. Kick, J.: Cyber exercise playbook (2014). URL https://www.mitre.org/news-insights/publication/cyber-exercise-playbook. Accessed: 15 September 2023
10. King, T.C., Aggarwal, N., Taddeo, M., Floridi, L.: Artificial intelligence crime: An interdisciplinary analysis of foreseeable threats and solutions. Science and engineering ethics **26**, 89–120 (2020)
11. Kokkonen, T., Puuska, S.: Blue team communication and reporting for enhancing situational awareness from white team perspective in cyber security exercises. In: O. Galinina, S. Andreev, S. Balandin, Y. Koucheryavy (eds.) Internet of Things, Smart Spaces, and Next Generation Networks and Systems, pp. 277–288. Springer International Publishing, Cham (2018)
12. Longbine, D.F.: Red Teaming: Past and Present. School of Advanced Military Studies, Fort Leavenworth, Kansas (2008)
13. McGowan, J., Straus, S., Moher, D., Langlois, E.V., O'Brien, K.K., Horsley, T., Aldcroft, A., Zarin, W., Garitty, C.M., Hempel, S., Lillie, E., Özge Tunçalp, Tricco, A.C.: Reporting scoping reviews—PRISMA ScR extension. Journal of Clinical Epidemiology **123**, 177–179 (2020). DOI 10.1016/j.jclinepi.2020.03.016. URL https://doi.org/10.1016%2Fj.jclinepi.2020.03.016
14. Munn, Z., Peters, M.D., Stern, C., Tufanaru, C., McArthur, A., Aromataris, E.: Systematic review or scoping review? guidance for authors when choosing between a systematic or scoping review approach. BMC medical research methodology **18**, 1–7 (2018)
15. NATO Standardization Office (NSO): Nato standard app-6, nato joint military symbology. Standard Edition D, Version 1, North Atlantic Treaty Organization (NATO) (2017)
16. Pistono, F., Yampolskiy, R.V.: Unethical research: how to create a malevolent artificial intelligence. In: Proceedings of Ethics for Artificial Intelligence Workshop (AI-Ethics-2016), pp. 1–7 (2016)
17. Renaud, K., Warkentin, M., Westerman, G.: From ChatGPT to HackGPT: Meeting the cybersecurity threat of generative AI. MIT Sloan Management Review (2023). Reprint #64428
18. Smith, J., Theisen, C., Barik, T.: A case study of software security red teams at Microsoft. In: 2020 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC), pp. 1–10. IEEE (2020). DOI 10.1109/VL/HCC50065.2020.9127203
19. Sommestad, T., Hallberg, J.: Cyber security exercises and competitions as a platform for cyber security experiments. In: A. Jøsang, B. Carlsson (eds.) Secure IT

Systems, pp. 47–60. Springer Berlin Heidelberg, Berlin, Heidelberg (2012). DOI 10.1007/978-3-642-34210-3_4
20. Truong, T.C., Diep, Q.B., Zelinka, I.: Artificial intelligence in the cyber domain: Offense and defense. Symmetry **12**(3), 410 (2020)
21. Wang, C., Chen, J., Yang, Y., Ma, X., Liu, J.: Poisoning attacks and countermeasures in intelligent networks: Status quo and prospects. Digital Communications and Networks **8**(2), 225–234 (2022)
22. Wang, W., Siau, K.: Artificial intelligence, machine learning, automation, robotics, future of work and future of humanity: A review and research agenda. Journal of Database Management **30**(1), 61–79 (2019). DOI 10.4018/JDM.2019010104
23. Ward, D., Wooderson, P.: Automotive Cybersecurity: An Introduction to ISO/SAE 21434, p. 106. SAE International (2021)
24. Wilhelmson, N., Svensson, T.: Handbook for planning, running and evaluating information technology and cyber security exercises. The Swedish National Defence College, Center for Asymmetric Threats Studies (CATS) (2014)
25. Yamin, M.M., Ullah, M., Ullah, H., Katt, B.: Weaponized AI for cyber attacks. Journal of Information Security and Applications **57**, 102,722 (2021)
26. Yuen, J.: Automated Cyber Red Teaming. DSTO Defence Science and Technology Organisation, Edinburgh, Australia (2015)
27. Zhou, W.C., Sun, S.L.: Red Teaming Strategy: Huawei's Organizational Learning and Resilience, pp. 299–317. Springer International Publishing, Cham (2020). DOI 10.1007/978-3-030-47579-6_13
28. Zouave, E., Bruce, M., Colde, K., Jaitner, M., Rodhe, I., Gustafsson, T.: Artificially intelligent cyberattacks. Swedish Defence Research Agency, FOI, Tech. Rep. FOI (2020)