Taylor & Francis
Taylor & Francis Group

Applied Artificial Intelligence
An International Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uaai20

The Emerging Threat of Ai-driven Cyber Attacks: A Review

Blessing Guembe, Ambrose Azeta, Sanjay Misra, Victor Chukwudi Osamor, Luis Fernandez-Sanz & Vera Pospelova

To cite this article: Blessing Guembe, Ambrose Azeta, Sanjay Misra, Victor Chukwudi Osamor, Luis Fernandez-Sanz & Vera Pospelova (2022) The Emerging Threat of Ai-driven Cyber Attacks: A Review, Applied Artificial Intelligence, 36:1, 2037254, DOI: 10.1080/08839514.2022.2037254

To link to this article: https://doi.org/10.1080/08839514.2022.2037254

Published online: 04 Mar 2022.

APPLIED ARTIFICIAL INTELLIGENCE
2022, VOL. 36, NO. 1, e2037254 (2409 pages)
https://doi.org/10.1080/08839514.2022.2037254

Taylor & Francis
Taylor & Francis Group

The Emerging Threat of Ai-driven Cyber Attacks: A Review

Blessing Guembe[a], Ambrose Azeta[b], Sanjay Misra[c], Victor Chukwudi Osamor[a], Luis Fernandez-Sanz[d], and Vera Pospelova[d]

[a]Department of Computer and Information Sciences, Covenant University, Ota, Ogun State, Nigeria;
[b]Department of Computer Science, Namibia University of Science and Technology, NAMIBIA;
[c]Department of Computer Science and Communication, Ostfold University College, Halden, Norway;
[d]Department of Computer Science, University of Alcala, Madrid, Spain

ABSTRACT

Cyberattacks are becoming more sophisticated and ubiquitous. Cybercriminals are inevitably adopting Artificial Intelligence (AI) techniques to evade the cyberspace and cause greater damages without being noticed. Researchers in cybersecurity domain have not researched the concept behind AI-powered cyberattacks enough to understand the level of sophistication this type of attack possesses. This paper aims to investigate the emerging threat of AI-powered cyberattacks and provide insights into malicious used of AI in cyberattacks. The study was performed through a three-step process by selecting only articles based on quality, exclusion, and inclusion criteria that focus on AI-driven cyberattacks. Searches in ACM, arXiv Blackhat, Scopus, Springer, MDPI, IEEE Xplore and other sources were executed to retrieve relevant articles. Out of the 936 papers that met our search criteria, a total of 46 articles were finally selected for this study. The result shows that 56% of the AI-Driven cyberattack technique identified was demonstrated in the access and penetration phase, 12% was demonstrated in exploitation, and command and control phase, respectively; 11% was demonstrated in the reconnaissance phase; 9% was demonstrated in the delivery phase of the cybersecurity kill chain. The findings in this study shows that existing cyber defence infrastructures will become inadequate to address the increasing speed, and complex decision logic of AI-driven attacks. Hence, organizations need to invest in AI cybersecurity infrastructures to combat these emerging threats.

ARTICLE HISTORY
Received 5 November 2021
Revised 26 January 2022
Accepted 28 January 2022

Introduction

Cyberattacks are pervasive and are often regarded as one of the most tactically significant risks confronting the world today (Dixon and Eagan 2019). Cybercrimes can engender disastrous financial losses and affect individuals and organizations as well. It is estimated that a data breach costs the United States around 8.19 million Dollars and 3.9 Million Dollars on average, and the annual effect on the global economy from cyberattack is approximately 400

CONTACT Sanjay Misra sanjay.misra@hiof.no Department of Computer Science and Communication, Ostfold University College, Norway

This article has been republished with minor changes. These changes do not impact the academic content of the article.

Billion Dollars (Fischer 2016; Kirat, Jang, and Stoecklin 2018). A Cyberattack is the intentional exploitation of computer systems, networks, and businesses. With increasingly sophisticated cybersecurity attacks, cybersecurity specialists are becoming incapable of addressing what has become the most significant threat climate ever before (Chakkaravarthy et al., 2018).

The sophistication of cyberattack techniques poses an existential danger to enterprises, essential services, and organization infrastructures, with the power to interrupt corporate operations, wipe away critical data, and create reputational damage. Today's current wave of attacks outwits and outpaces humans and even includes Artificial Intelligence (AI). Cybercriminals will be able to direct targeted attacks at unprecedented speed and scale while avoiding traditional, rule-based detection measures thanks to what's known as "offensive AI" (DarkTrace, 2021). A new generation of cybercriminals has emerged, one that is both subtle and secretive, which will influence the future of cybersecurity. The new generation of cyber threats will be smarter and capable of acting independently with the help of AI. Future cyberattack methods will be able to be aware of their surroundings and make informed decisions based on the target environment. The potential of AI to learn and adapt will usher in a new era of scalable, custom-made, and human-like assaults (Thanh and Zelinka 2019).

Cybercriminals today have many sophisticated AI-driven cyberattacks methods by which they can create problems for the government, organizations, businesses, and individuals. Existing cybersecurity tools are no longer viable against this advanced cyber-weaponry (Thanh and Zelinka 2019). The negative use of AI to compromise digital security is known as AI-driven cyberattack, in which cybercriminals can train robots to socially engineer targets at human or superhuman levels of performance (Brundage et al. 2018). AI-assisted attacks will be able to adapt to the environment it infects. Learning from contextual data to emulate trustworthy features of cyberspace or target weak points it discovers (DarkTrace, 2020). AI-driven cyberattacks are not a far-fetched future threat. The necessary tools and building blocks for launching an offensive AI-driven cyberattack already exist (Dixon and Eagan 2019). Recent advancements in AI have influenced the tremendous growth in automation and innovation. Although these AI systems have many benefits, they can also be used maliciously. AI-driven cyberattacks have been on the rise in recent years. Even as many organizations transit to more secure cyberspace, such as cloud storage services, their resources remain vulnerable to cyber criminals (Bocetta 2020). In general, AI-driven cyberattacks will only worsen, then it will be almost impossible for traditional cybersecurity tools to detect them. It is simply a question of machine efficiency vs. human effort. The complexity and size of this growing trend are submerging cybersecurity teams. In contrast, the advanced and qualified cybersecurity specialists necessary to counter this threat successfully are increasingly expensive and difficult to find. The consequences of these emerging AI-driven attack techniques

could be life-threatening and highly destructive. These subtle attacks undermine trust in organizations by undermining data security and integrity, potentially resulting in systemic failures (Cabaj, Kotulski, Księżopolski & Mazurczyk, 2018; Hamadah & Aqel, 2020). An AI-driven cyberattack will harness a multitude of cyberspace and computer resources well beyond what a human could enlist, resulting in an attack that is faster, more unpredictable, more sophisticated than even the strongest cybersecurity team can respond against. (Bocetta 2020). As AI becomes a more powerful tool in the hands of malicious actors, cybersecurity researchers, practitioners, and governments will need to respond with more inventive solutions to successfully safeguard cyberspace from malicious actors (Hamadah & Aqel, 2020). AI-driven attacks employ sophisticated obfuscating algorithms and frequently change identifiable characteristics that could reach a level of adaptability that renders it virtually undetectable to both behavioral and signature-based antivirus tools (Babuta, Oswald & Janjeva, 2020). High-security risks and elusive attacks in benign carrier applications, such as DeepLocker, have demonstrated the intentional use of AI for negative objectives. Attackers are constantly improving and changing their attack strategy, with a particular focus on the use of AI-based techniques in the attack process, known as AI-driven cyberattacks, which can be used in collaboration with traditional cyberattack techniques to cause more significant damage in cyberspace while remaining undetected (Kaloudi and Li 2020; Thanh and Zelinka 2019; Usman et al. 2020).

AI-driven cyberattack techniques will have the capacity to adapt to the surroundings where it executes. They can exploit the vulnerabilities or masquerade as trusted system attributes by learning from contextual data or information. The longer the attack exists in the host, the more they integrate and become independent of its targets, environments, and countermeasures against cybersecurity defense infrastructures (Thanh and Zelinka 2019). The consequences of these emerging AI-driven attack techniques could be life-threatening and highly destructive. Hence, this study investigates the emerging threat of AI-driven attacks and reviews the negative impacts of this sophisticated cyber weaponry in cyberspace.

The paper is divided into five parts. The mechanism for offering the review process is presented in the next section. Section 3 contains the results. In section 4, the findings are presented, and in section 5, a conclusion is formed.

Research Questions and Review Process

This study investigates the emerging threat of AI-driven cyberattacks and the techniques utilized by cybercriminals to carry out AI-driven cyberattacks, focusing on literature that addresses the research questions. Table 1 illustrates the three research questions as well as the justification, which also establishes a foundation for identifying the studies and formulation of our search criteria.

Table 1. Research Questions.

Research questions	Rationale
RQ1: What are the current and emerging AI-driven techniques malicious attackers utilize to carry out cybercrime?	To identify the existing AI techniques malicious actors utilize to cause more significant damage in cyberspace without being noticed.
RQ2: What is the difference between traditional targeted cyberattacks and AI-driven cyberattacks?	Identify the difference between traditional targeted cyberattacks and AI-driven attacks and identify their various logical components.
RQ3: What are the impacts of AI-driven cyberattacks?	To identify the effects of AI-driven attacks and how these techniques can be enhanced in the future to conceal sophisticated cyber threats.

Research Methodology and Eligibility Criteria

This study's systematic literature review methodology was based on Prisma International Standards (Moher, Liberati, Tetzlaff & Altman, 2010). This research aimed to find aset of studies on the topic of AI-driven cyberattacks that were relevant. To do so, the authors outlined the research questions offered in this study and explained the reasoning behind each one. The authors also discuss the criteria for selecting the required literature and the search method for retrieving relevant data and published publications.

Search Criteria and Identification of Studies

The following search parameters were used to find relevant literature for this study:

(1) Make a list of keywords from the research questions.
(2) Identify keywords in relevant literature;
(3) Recognize distinct keyword synonyms and spellings;
(4) To relate primary keywords and concepts using the Boolean operators "AND" and "OR."

The search keywords are developed from the research questions in Table 1. The output of the search string used for searching relevant literature is as follows: ("AI-driven cyberattacks" OR "AI-driven attack techniques" OR "malicious use of AI") AND ("AI-powered cyberattacks" OR "AI-based cyber-attack" OR "Artificial intelligence in cyberattack" OR "Impact of AI-Driven attack").

Three factors were used to apply the eligibility criteria: inclusion, exclusivity, and quality criteria. These criteria were used to extract the literature from the search results.

Exclusive Criteria

The following exclusive criteria were used to evaluate the retrieved literature:

EC1:Non-discussion of the research questions in the literature.

EC2:Articles on the same subject.

EC3:The same articles from different databases.

EC4:Articles that do not discuss AI-driven cyberattacks.

Inclusion Criteria

The retrieved literature was evaluated with the following inclusive criteria:

IC1:Literature that is relevant to AI-driven cyberattacks.

IC2:Methodologies, Journals, Conference and White Papers that addressed AI-driven attacks.

IC3:Papers address AI-driven attacks techniques such as deep learning, bio-inspired swarm intelligence, etc.

Quality Criteria

The retrieved papers were screened based on the following quality criteria:

QC1:Do the papers answer the majority of the research questions?

QC2:Is it possible to describe the AI techniques used for executing AI-Driven attacks?

QC3:Do the studies provide answers to any of the research questions?

QC4:Do the papers adequately disclose the research methodology?

Selecting Procedure

The key search criteria for this study were ACM, arXiv Blackhat, Scopus, Springer, MDPI, and IEEE research databases. The PRISMA flowchart, depicted in Figure 1, depicts the systematic review process and selection of relevant papers at various phases. Following the presentation of the sources, criteria, and methodology for selecting relevant publications, a quantitative evaluation was demonstrated to discover new methodologies, measures, or contributions offered by researchers in the study domain.

(1) Stage 1 (Extracting Information): This is based on information extraction; a comprehensive search was conducted on nine electronic databases, yielding 936 article outlines, which served as a pool of potential articles for subsequent selection, as shown in Table 2.

(2) Stage 2 (Screening): A total of 936 potential articles were found based on Table 2. There are 417 duplicate papers among the 936 research articles because publications were found in more than one online resource. The screening process was then undertaken based on the title of the articles' irrelevancy, and 309 articles were deemed unsuitable for this study.

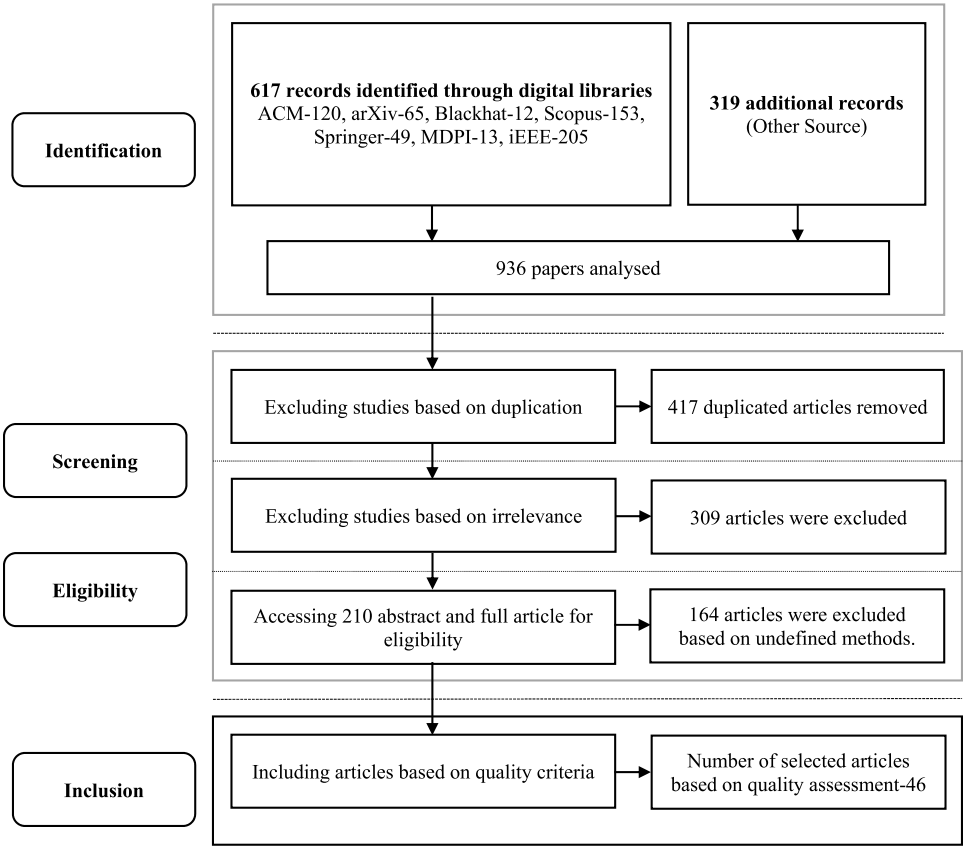

Figure 1. PRISMA flowchart illustrating the systematic review process and article selection at various stages.

Table 2. The number of publications found in online databases.

S/N	Database	No. of articles
1	ACM	120
2	arXiv	65
3	Blackhat	12
4	Others	319
5	Scopus	153
6	Springer	49
7	MDPI	13
8	IEEE	205

(3) Stage 3 (Eligibility): An article's relevance and quality cannot be determined solely by its abstract and title. As a result, complete text-based selection criteria were used to extract relevant articles for the study. A total of 210 papers were reviewed for eligibility, with 164 being eliminated due to imprecise technique or ambiguity.

(4) Stage 4 (Inclusion): The authors conducted a quality assessment based on the research of the above topics for the remaining papers. After resolving issues with article selection, a total of 46 primary papers were examined for this investigation.

Search Strategy

This section outlines and examines the various search strategies used to identify the 46 papers for this study. The forty-six articles were published as follows: fifteen papers for journals, fourteen papers from conference proceedings, eight articles in workshops, six from symposiums, two papers from white papers, and one article from scholarly work, as shown in Figure 2. The search approach was divided into four classes in this study to explore all contributions made by past researchers in this research domain. Techniques, status, source, and attack strategy are the four search techniques used in this study.

Sources

In this study, data was extracted from seven digital databases. ACM, arXiv Blackhat, Scopus, Springer, MDPI, and IEEE Xplore are available digital libraries. Published conference proceedings, journal papers, workshops, symposiums, and scholarly work were searched using titles, abstracts, and keywords.

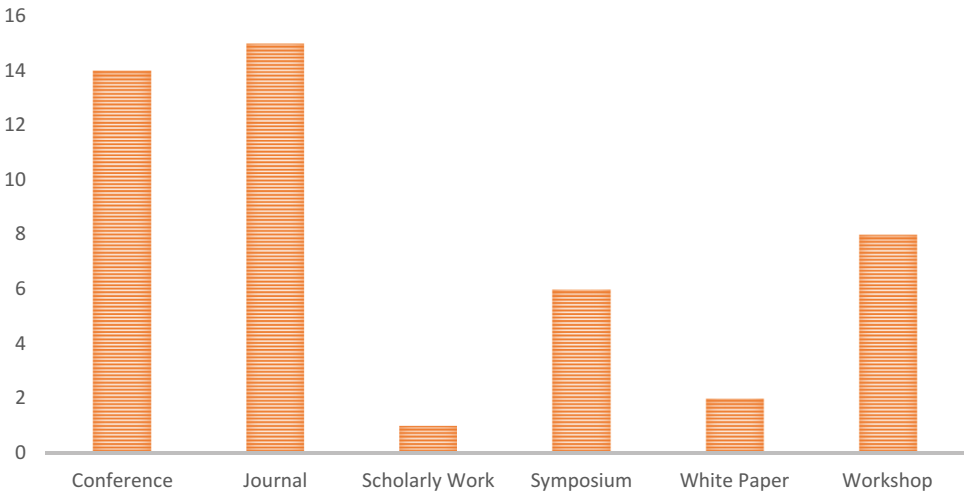

Figure 2. Number of collated studies.

Attack Strategy

In this paper, the AI-driven attack strategies identified in the forty-six selected papers are deep learning, bio-inspired computation and swarm intelligence, and fuzzy model. A large proportion of the selected papers are based on deep learning strategy, as described in Section 4.

Types of Attacks

On the basis of the 46 selected papers, this section identifies nineteen uses cases of offensive AI in six stages of the cybersecurity kill chain, as shown in Figure 3. In the access and penetration phase (AI-aided attack), six types of AI-driven attacks were identified, four types of AI-driven attacks were identified in the access reconnaissance stage (AI-targeted attack), four types of AI, three types of AI-driven attacks were identified in the exploitation stage (AI-automated attack), two types of AI-driven attacks were also identified in the delivery stage (AI-concealment attack) and C2 stage (AI-multi-layered attack) respectively. In contrast, one type of AI-driven attack was identified in action on objectives stage (AI-malware attack), as shown in Figure 3.

Figure 4 shows that the access and penetration stage has the most publications (6), followed by the reconnaissance stage (4), the exploitation stage has three publications, and the delivery and C2 stages have two. In contrast, the action on objectives stage has the least publication (1).

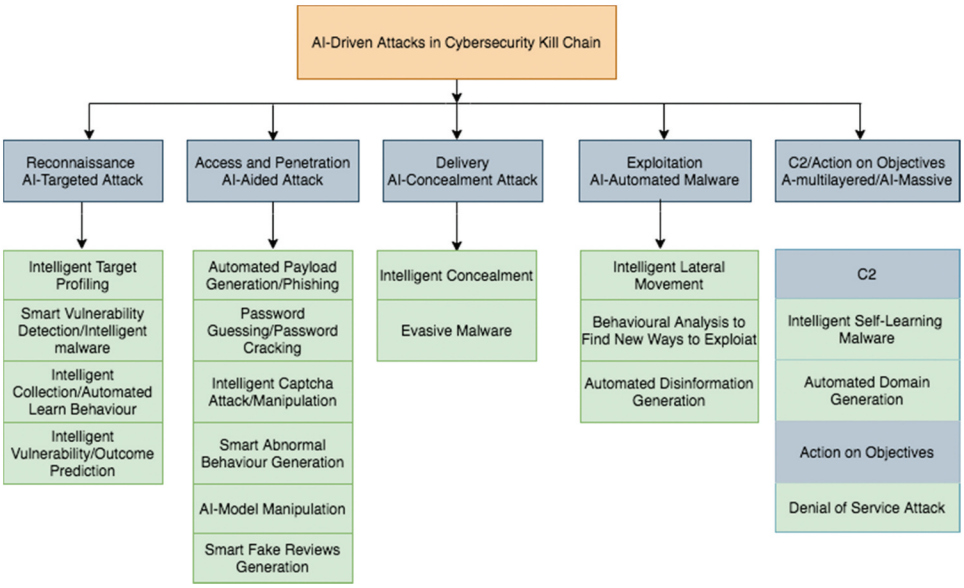

Figure 3. Modified Cybersecurity Kill Chain for AI-Driven Attack (Kaloudi and Li 2020).

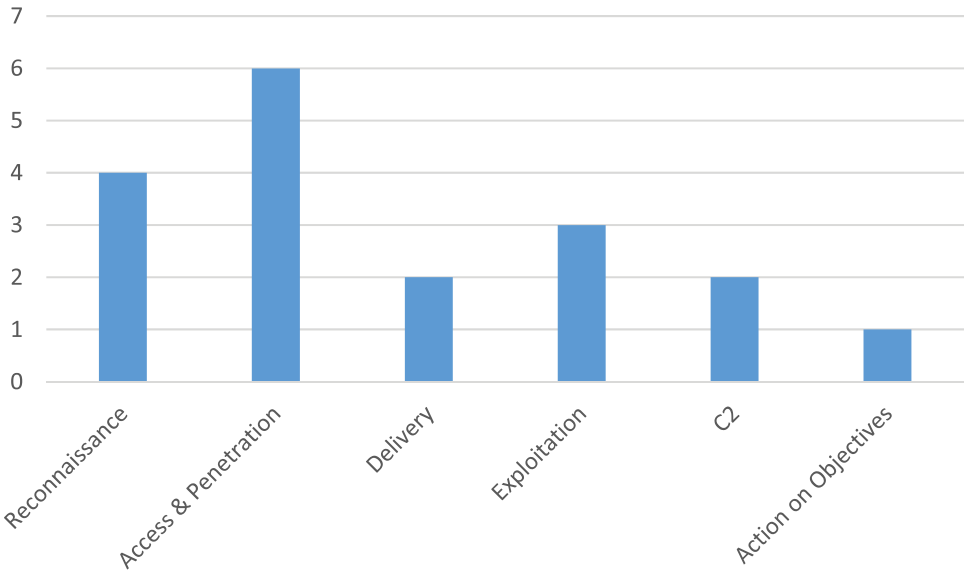

Figure 4. Offensive AI Techniques Cyberattacks in the Modified Cybersecurity Kill Chain.

Techniques

The selected studies demonstrated how malicious actors could utilize AI techniques to execute vulnerability prediction, End-to-End (E2E) spear-phishing, and intelligent target profiling/ intelligence collection in the cybersecurity kill chain reconnaissance phase, as discussed in the cybersecurity kill chain subsection 3.21. The selected studies also identified nineteen (19) AI techniques that malicious actors can utilize to execute attacks in the access and penetration phase of the cybersecurity kill chain. CNN has the highest appearance (five) AI techniques utilized by the authors to demonstrate access and penetration attacks. GAN and RNN were used two times respectively to demonstrate access and penetration attacks. In contrast, LSTM, SVC, SVM, cycle-GAN, TOD+CNN, RF, MP, GBRT, and KNN were demonstrated one time in malicious attacks as discussed in subsection 3.2.2. The selected articles identified three AI techniques that malicious actors can utilize to execute attacks in the delivery stage of the cybersecurity kill chain. Two of the studies identified GAN for intelligent concealment to generate adversarial malware and undetectable malware URLs. One of the studies utilized LSTM to generate automated malicious evasive payloads. At the same time, one study demonstrated how Malicious actors could utilize DNNs to conceal malicious intent and activate it when it gets to its specific target (Kirat, Jang, and Stoecklin 2018), as discussed in subsection 3.2.3. From the selected articles, three (3) AI techniques were utilized to demonstrate behavioral analysis to find new ways to exploit targeted infrastructures and automated disinformation generation in the delivery phase of the cybersecurity kill chain, as discussed in subsection

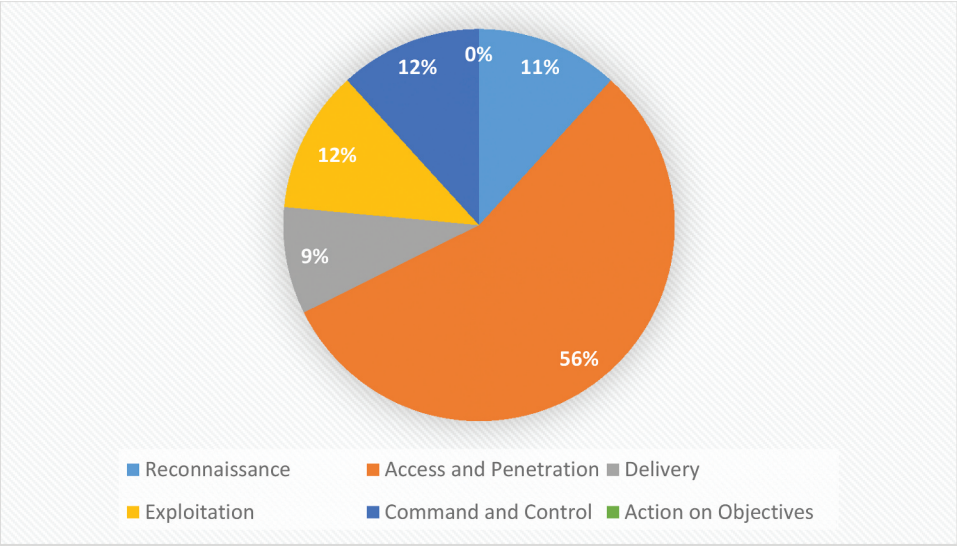

Figure 5. Identified AI-Driven Cyberattack Techniques.

3.2.4. One of the selected studies demonstrated how NNs and Reinforcement Learning (RL) could be utilized to execute behavioral analysis to exploit vulnerabilities in web-based applications in other to bypass web-based application authentication systems. K-means clustering was also utilized to demonstrate how AI-driven self-learning malware can successfully exploit vulnerabilities in security detection systems and act as though they were unintentional failures on computer applications by exploiting and compromising sensitive environmental control infrastructures in the exploitation phase of the cybersecurity kill chain. Markov chains and LTSM were utilized to execute automated machine-generated content disinformation by implementing an end-to-end spear-phishing technique to generate personalized content for high target users on Twitter. Two types of AI-driven cyberattacks were identified: intelligent self-learning malware and automated domain generation, as discussed in subsection 3.2.5. Four types of AI techniques identified malicious actors could utilize to execute AI-driven self-learning malware and automated domain generation attack in the command and control of the cybersecurity kill chain, as discussed in subsection 3.2.5. Two of the selected studies utilized K-means clustering, Gaussian distribution, and DNNs to demonstrate intelligent self-learning malware attacks, as discussed in subsection 3.2.5. 56% of the AI-Driven cyberattack technique were identified in the access and penetration phase of the cybersecurity kill chain, 12% AI techniques were identified in the exploitation and command and control phase, respectively, 11% AI techniques were identified in the reconnaissance phase, 9% AI

Table 3. Status of existing literature with respective narration.

S/N	State	Narration
1	Implemented	This refers to the category of studies that designed a proof of concept to demonstrate AI-driven attacks.
2	Proposed	This category of studies is based on new techniques or methods without any proof of concept and evaluation.
3	Implemented and evaluated	This category of studies refers to those that designed a proof of concept demonstrating AI-driven cyberattacks and evaluated the proof of concept based on performance metrics.

Figure 6. Generated keywords from titles of selected articles.

technique were identified in the delivery phase, and no AI technique was demonstrated in action on the objective stage of the cybersecurity kill chain to execute offensive AI attacks as shown in Figure 5.

Status

The literature chosen is grouped into three categories. This is accomplished as illustrated in Table 3.

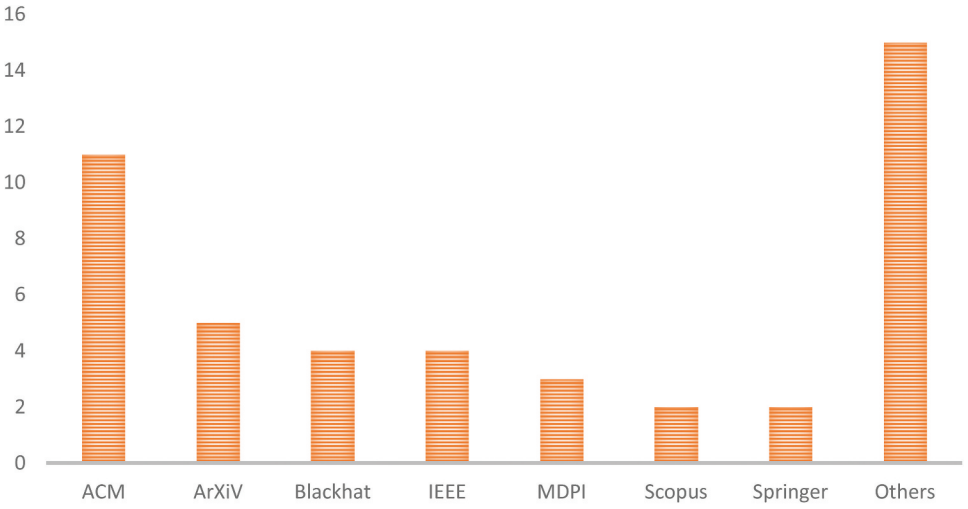

Figure 7. Average numbers of relevant articles.

Result Obtained

The results of the four search methodologies were analyzed in detail in this section, which discussed the study findings. In several subsections, the study presented a full discussion of the findings in relation to the outline study topics, along with a concise interpretation of our findings. The results of a word cloud analysis employing titles of selected articles on the orange machine learning integrated environment are shown in Figure 6, with 'Learning' being the most common occurrence, followed by 'Artificial,' 'Machine,' 'attack,' 'attacks,' and 'cybersecurity.'

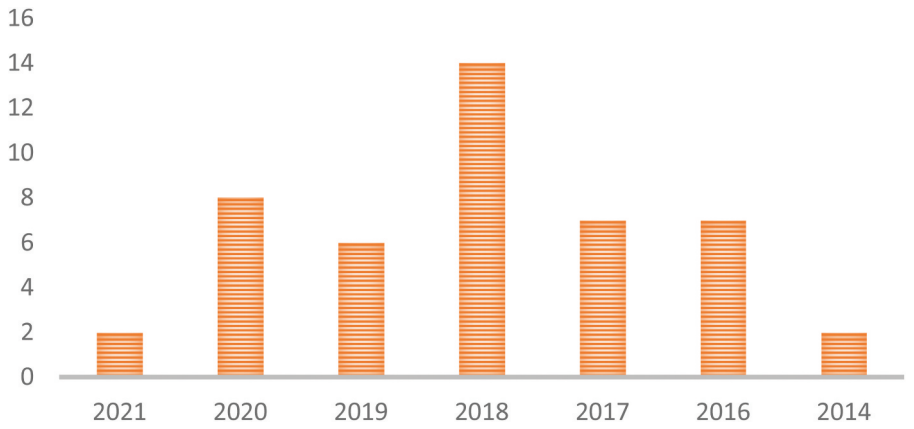

Figure 8. Publication Year.

Search Strategy 1: Source

The initial search exercise was carried out using a hierarchical search method to find related articles on AI-driven cyberattacks using the article title and keywords before designing a final search strategy.

The following databases were utilized to find related literature for publications published between 2014 and 2021: ACM, arXiv Blackhat, MDPI, Scopus, Springer, and IEEE Xplore. The retrieved findings for relevant article sources are shown in Figure 7, while the number of relevant publications released during the research year is shown in Figure 8.

Figure 5 shows that of the final 46 papers, ACM had the most relevant publications with eleven (11), followed by arXiv with five (5), Blackhat and IEEE have four (4) publications respectively, Scopus and Springer have two (publications) respectively, while other fifteen (15) relevant publications were retrieved from other sources.

Figure 6 demonstrates that 2018 had the most relevant papers (14), whereas 2014 had the least publications (2) respectively. The threat of AI-driven cyberattacks grows despite ongoing research attempts to understand and combat these advanced cyber weapons.

Figure 9 depicts the AI techniques used by the selected studies to demonstrate the malicious use of AI in cyberattacks in the access and penetration stage of the modified cybersecurity kill chain.

Figure 10 depicts the AI techniques used by the selected authors to demonstrate the malicious use of AI in the delivery stage of the modified cybersecurity kill chain. The results indicate that GAN has the most publications (2), while DNNs and LSTM have one publication, respectively.

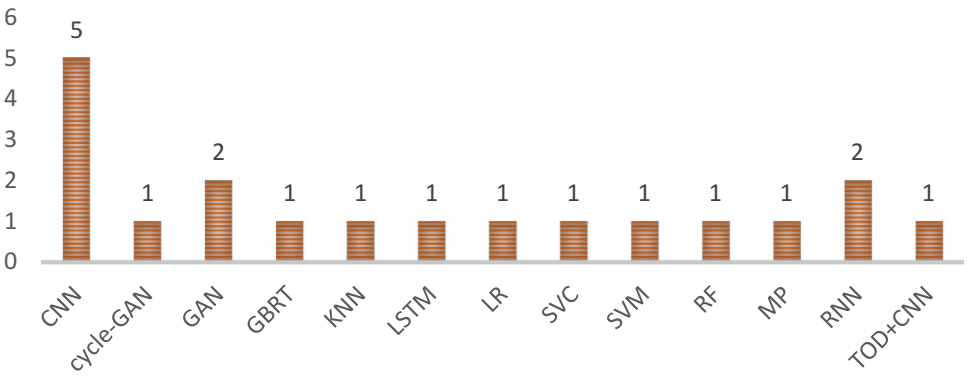

Figure 9. AI Techniques in the Access and Penetration Attack Stage.

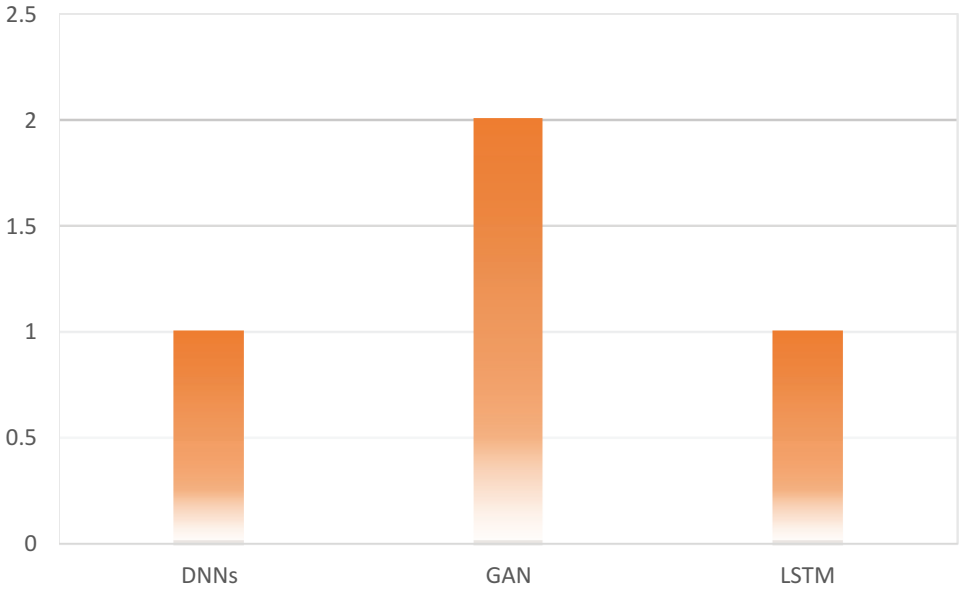

Figure 10. AI Techniques in the Delivery Stage.

Search Strategy 2: Techniques

The outcome of this segment is broken down into six stages of the cybersecurity kill chain, which include reconnaissance (AI-targeted attack), access and penetration (AI-aided attack), delivery (AI-concealment attack), exploitation (AI-automated malware), command on control (AI-multi-layered attack), action on objectives (AI-massive attack).

Reconnaissance Stage

Three types of AI techniques were identified in the reconnaissance stage of the cybersecurity kill chain. The selected studies demonstrated how malicious actors could utilize Markov chains/LTSM, NNs, DNNs to execute vulnerability prediction, End-to-End (E2E) spear-phishing, and intelligent target profiling/intelligent collection, respectively. Table 4 summarizes the AI techniques utilized to execute AI attacks in the reconnaissance stage.

Access and Penetration Stage

This study identified six (6) AI-driven attacks in the access and penetration phase. They include; password guessing/password cracking (brute-force attack), intelligent captcha/manipulation, smart abnormal behavioral generation, AI model manipulation, and smart fake reviews generation. The study also identified nineteen (19) AI techniques that can be utilized

Table 4. Offensive Utilization of AI Techniques in the Reconnaissance Stage of Cybersecurity Kill Chain.

Authors	Attack	Class of Attack	Attack Goal	Technique
Seymour and Tully (2016)	E2E spear phishing	Intelligent Malware	An automated end-to-end spear-phishing strategy that includes identifying high-value targets and propagating personalized machine-generated information on Twitter.	Markov chains and LTSM
Dheap (2017)	Vulnerability Prediction	Outcome Prediction	Boost malicious actors' confidence to seek risker and high-value outcomes to overpower state-of-the-art cybersecurity infrastructures.	NNs
Kirat, Jang, and Stoecklin (2018).	DeepLocker	Intelligent Target Profiling/ Intelligent Collection	Hide payload without being detected in video conferencing application.	DNNs

Table 5. Offensive Utilization of AI Techniques in the Access and Penetration Stage of Cybersecurity Kill Chain.

Authors	Attack	Class of Attack	Attack Goal	Technique
Bahnsen et al. (2018).	DeepPhish: Simulating Malicious AI.	Automated Payload/ Phishing.	To prevent AI cyberattack detection model and execute more effective phishing attacks.	LSTM.
Hitaj et al. (2019).	PassGAN.	Password Guessing.	Guess password based on learning the distribution of actual password leaks.	GAN.
Trieu and Yang (2018).	Intelligent password brute force attack.	Password Cracking	To obtain previous password sequences, in order to create new passwords by guessing one character at a time.	RNN
Lee and Yim (2020).	Offensive password authentication technique.	Password Guessing.	Predict and steal users' actual passwords based on keyboard strokes.	LR, SVM, SVC, RF, KNN, GBRT, MP
Bursztein et al. (2014).	Single Step Captcha Solver.	Intelligent Captcha Attack.	To Combine segmentation and recognition issues to attack captcha in a single phase.	CNN.
Gao et al. (2017)	Breaking text-based Captchas.	Intelligent Captcha Attack.	Four deep learning models with 2, 3, 5, and 6 convolutional layers as recognition engines to attack text-based captchas.	CNN.
Ye et al. (2018)	Text-based Captcha solver.	Intelligent Captcha Attack.	To generate synthetic captchas and then fine-tune the base solver on a limited selection of real Captchas using transfer learning.	GAN
Gao et al. (2017).	Attacking two-layer captcha.	Intelligent Captcha Attack.	To break two-layer captchas with an enhanced LeNet-5 and a radical CNN model as a recognition engine.	CNN.
Yu and Darling (2019).	AI-based Captcha solver.	Intelligent Captcha Attack.	To determine which character was contained in a segmented sample in order to crack captcha.	TOD+CNN
Noury and Rezaei (2020).	Captcha solver for vulnerability assessment	Intelligent Captcha Attack.	To explore the flaws and weaknesses of existing Captcha generator systems.	CNN.
Chen et al. (2018).	Hollow Captcha solver.	Intelligent Captcha Attack.	To improve attack accuracy and reduce the attack time, use precise filling and nonredundant merging.	CNN.
Li et al. (2021).	End-to-end attack on text-based Captchas.	Intelligent Captcha Attack.	Using Captcha synthesizers based on the cycle-GAN, create some false samples.	cycle-GAN
Yao et al. (2017).		Smart Fake Review Generation.	To generate fake smart reviews.	RNN

by malicious actors to execute access and penetration attacks. Table 5 summarizes the AI techniques utilized to execute AI attacks in the access and penetration stage.

Delivery Stage

In the delivery phase, two types of AI-driven cyberattacks were identified: intelligent concealment and evasive malware. The selected studies identified three types of AI techniques; malicious actors could utilize that to execute AI-driven concealment and evasive attack, as illustrated in Table 6 and Figure 8. The study identified three AI techniques that malicious actors can utilize to execute attacks in the delivery stage. Table 6 summarizes the AI techniques utilized to execute AI attacks in the delivery stage.

Exploitation Stage

The exploitation phase involves gaining authorized access to computer applications and resources, and after gaining access to the target, malicious actors can utilize AI techniques to execute complex attacks that are difficult to detect using NNs and DNNs. Table 7 summarizes the AI techniques utilized to execute AI attacks in the exploitation stage.

Command and Control Stage

In the command and control (C2) phase, two types of AI-driven cyberattacks were identified: intelligent self-learning malware and automated domain generation. Table 8 summarizes the AI techniques utilized to execute AI attacks in the C2 stage.

Table 6. Offensive Utilization of AI Techniques in the Delivery Stage of Cybersecurity Kill Chain.

Authors	Attack	Class of Attack	Attack Goal	Technique
Bahnsen et al. (2018).	Malicious Payload.	Intelligent Concealment.	Automated generation of undetected phishing URLs.	LSTM
Hu and Tan (2021).	Adversarial malware generation.	Intelligent Concealment.	Generating undetectable adversarial malware to bypass machine learning black-box cyber threat detection systems	GAN
Anderson, Woodbridge, and Filar (2016).	Undetectable malware URL.	Intelligent Concealment.	GAN-based automatic generation of undetectable malware URL that learns to bypass DNNs-based malware detection system.	GAN
Kirat, Jang, and Stoecklin (2018).	DeepLocker	Evasive Malware	Conceal its attack and only activate it for specific targets.	DNNs

Table 7. Offensive Utilization of AI Techniques in the Exploitation Stage of Cybersecurity Kill Chain.

Authors	Attack	Class of Attack	Attack Goal	Technique
Petro and Morris (2017).	DeepHack: Open source Hacking AI	Behavioral analysis to exploit vulnerabilities.	Breaking into web applications.	NNs, Reinforcement Learning (RL).
Chung, Kalbarczyk, and Iyer (2019)	Self-Learning Malware.	Behavioral analysis to exploit vulnerabilities.	A malicious attack that exploits and compromises environmental control systems while masquerading as an unintentional failure on computer infrastructures.	K-means clustering
Seymour and Tully (2016)	Machine generated spear-phishing.	Automated disinformation generation	To generate personalized content for high-value targets on Twitter.	Markov chains and LTSM

Table 8. Offensive Utilization of AI Techniques in the C2 of Cybersecurity Kill Chain.

Authors	Attack	Class of Attack	Attack Goal	Technique
Chung, Kalbarczyk, and Iyer (2019)	To execute attack planning phase.	Intelligent self-learning malware	To attack computers at a supercomputer facility without the attacker's knowledge by interfering with the cyber-physical systems.	K-means clustering, and Gaussian distribution
Anderson, Woodbridge, and Filar (2016)	DGA classifier to analyze infected hosts.	Automated domain Generation.	To assess successful DNS queries made by infected hosts and assign scores based on values derived from training datasets.	GAN
Kirat, Jang, and Stoecklin (2018)	DeepLocker for self-learning malware.	Intelligent self-learning malware	To establish a self-learning attack in the C2 channel.	DNN

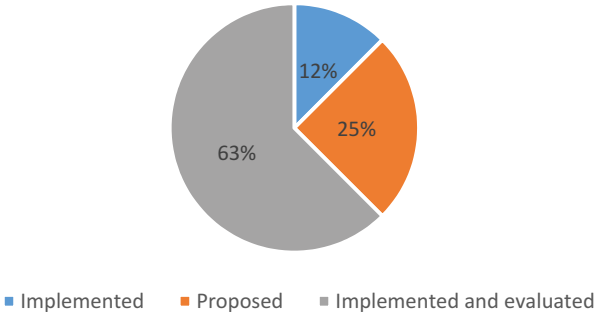

Figure 11. Status of existing AI-driven cyberattack tools.

Search Strategy 3: Status

The current state of the AI-driven attack tool was assessed using three categories, and the results of the analysis for the selected publications are shown in Figure 11. Based on existing AI-driven cyberattack technologies, the analysis result for the 46 articles evaluated in this study shows that 63% of the existing AI-driven cyberattack tools are implemented and evaluated, 25% are proposed only, and 12% of the existing AI-driven cyberattack tools are implemented without evaluation. As a result of the current state of the study, the majority of existing AI-driven cyberattack tools are based on implementation and evaluation.

Discussion on Research Questions

This section reviews the core concepts that make up this study, such as the current and emerging AI-Driven cyberattack techniques, weaponization of machine learning and deep learning techniques in cyberattacks, and types of AI-Driven attacks in the cybersecurity kill chain and existing AI-Driven attacks. Also discussed in the section are the result and findings of this study.

RQ1:Current and Emerging AI-Driven Cyberattack Techniques

The advancement of cyberattack tools and recent techniques are shaping and expanding the cyberattack domain, which opens up cyberspace to a wide range of sophisticated cyber weaponry with many powerful negative effects (Kaloudi and Li 2020). Brundage et al. (2018) established a scenario that notifies cybersecurity researchers and industry about the malevolent utilization of AI by embedding some hypothetical concepts within digital, physical and political security domains. Researchers have established a few concepts that showed the potential of an automatic exploit generation in state-of-the-art applications.

Malicious actors are utilizing fuzzy models to develop a next-generation malware capable of learning from its environment, continuously updating itself with new variants, and infecting vulnerable and sensitive computer infrastructures without being noticed (Kaloudi and Li 2020). Malicious actors can utilize these concepts to deploy a new type of sophisticated and stealthy cyber weaponries.

AI-Driven Attacks in the Reconnaissance Stage of the Cybersecurity Kill Chain

Malicious actors can use AI techniques to improve reconnaissance to study normal behavior and operations about a cybersecurity defense mechanism, computer infrastructures, and devices (Kaloudi and Li 2020). In this case, a malicious actor can obtain structural, operational, and topological data about the user's devices, network flows, and network infrastructure to identify a critical relationship with the intended targets. Malicious actors may be able to use AI technology to detect patterns of targeted attacks in massive volumes of data. A reconnaissance attack, also classified as AI-targeted, depends on a well-prepared planning phase to execute its attack. AI's capability to interpret, discover and comprehend patterns in large amounts of the dataset can be utilized to provide in-depth analysis and to develop targeted exploration processes by overcoming human limitations (Kaloudi and Li 2020). The authors identified four AI-driven threat use cases in the reconnaissance phase of AI-targeted attacks; they include intelligent target profiling, clever vulnerability detection/intelligent malware, intelligent collection/automated learn behavior, and intelligent vulnerability/outcome prediction.

Intelligent Target Profiling. AI has already been shown to have an impact on the ability to profile the use of information and communication technology. Bilal et al. (2019) presented a taxonomy of profiling approaches as well as the AI algorithms that enable them. The authors pointed out that there are two forms of profiling: individual and group profiling, and that fuzzy logic ontology, machine learning, and convolutional neural networks are the most commonly used AI methodologies. With the advancement of AI techniques, cyberattack targets can be profiled based on their social media activity and public social media profiles. To maximize persuasive potential, AI systems may allow groups to target precisely the correct message at precisely the right time (Brundage et al. 2018). Malicious actors can utilize AI techniques to improve the chances of profiling their targets (Kirat, Jang, and Stoecklin 2018). Malicious actors can utilize DNNs and NNs to classify and profile targets (Dheap 2017). It is possible to draw conclusions from research prototypes like SNAP_R that both the technology readiness level and the chance of malicious end applications for executing intelligent profiling are high (Seymour and Tully 2016)

Intelligent Collection. Intelligence collection is a type of reconnaissance that aids in the planning and formulation of cyberattack policies. The AI techniques utilized to execute this type of attack include NLP and DNNs. These analyses should be classified as dual-use because they automate the collection of generic information on specific types of attacks and specific features that affect risk (both defensive and offensive) (Dheap 2017; Kirat, Jang, and Stoecklin 2018).

Intelligent Malware. By infiltrating the environmental control systems, intelligent malware §can initiate indirect cyber weaponries that pretend to be unintentional failures on computing infrastructure (Chung, Kalbarczyk, and Iyer 2019). Malicious actors can use a highly automated end-to-end spear-phishing technique that involves identifying high-priority targets and automatically disseminating personalized machine-generated information (Seymour and Tully 2016).

Outcome Prediction. AI techniques can examine current and previous occurrences in order to predict the outcomes of planned activities in the future. Cyber-related evaluation and simulation development techniques could be critical steps toward more advanced AI prediction models. For cybercriminals, offensive AI could increase their confidence in pursuing high-risk, high-value outcomes in order to defeat state-of-the-art cybersecurity systems (Dheap 2017).

AI-Driven Attacks in the Access and Penetration Stage of the Cybersecurity Kill Chain

This phase of cyberattack is also known as an AI-aided attack. This study identified six (6) AI-driven attacks in the access and penetration phase. They include; password guessing/password cracking (brute-force attack), intelligent captcha/manipulation, smart abnormal behavioral generation, AI model manipulation, and smart fake reviews generation.

Automated Payload Generation/Phishing. Malicious actors are capable of weaponizing machine learning algorithms to improve phishing attacks and make them invisible by cybersecurity detection systems, as demonstrated by Bahnsen et al. (2018) in DeepPhish: Simulating Malicious AI. DeepPhish is an AI algorithm that learns patterns from the most effective phishing URLs in the past to generate new synthetic phishing URLs. The objective is to create more effective phishing URLs to avoid AI detection and conduct more effective phishing attacks. To create phishing URLs in the past, attackers employed randomly generated segments. By utilizing an LSTM model to create a phishing URL classifier and produce new effective synthetic phishing URLs, the authors demonstrated the effectiveness of phishing attacks' by improving the

efficiency and success rate. The authors claimed that by training DeepPhish on two different threat actors raised the attack's effective rate from 0.69% to 20.9% and from 4.91% to 36.28%.

AI-Driven Password Guessing/Password Cracking. Three types of AI-driven password attacks were identified; password brute-force attack, password guessing, and password stealing. A deep learning model for password guessing was proposed by Hitaj et al. (2019). The authors evolved an automated password guessing technique based on GAN by learning the distribution from actual password breaches. Brute force, which entails testing all possible character combinations exhaustively; a dictionary, which entails using a list of likely words and previous password leaks in the hopes of correctly guessing; and rule-based approaches, which entail defining generation rules for possible password transformations such as concatenation.

Hitaj et al. (2019) evolved a model for correctly training a GAN such that tailored samples can be generated from the training set. GAN is used to automatically create passwords in the following way: The GAN is made up of a generating DNN (G) and a discriminative DNN (D) (D). There is also a training dataset using actual password samples, which are a collection of leaked passwords. A noise vector was used to train the generator G, which represents a random probability distribution and generates a sequence of vectors known as false password samples. The real and false samples are fed into the discriminator D, which subsequently learns to tell the difference between the two. When attempting to understand the original distribution of true password leaks, G compel D to disclose data to G. The rockyou dataset, which is an industry-standard password list, was used to train PassGAN, which achieved an effective result of both guessing new unique passwords and mimicking the distribution of rockyou dataset. PassGAN was able to correctly match 10,478,322 (24.2%) out of the 43,354,871 unique passwords from the LinkedIn data breach. GAN was never exposed to any of the LinkedIn datasets, but it was nevertheless able to produce meaningful, unique passwords based on the rockyou words. PassGAN, in combination with HashCat, was able to predict between 51 and 73% of unique passwords more accurately than HashCat alone.

Lee and Yim (2020) implemented a K-Nearest Neighbors (KNN), logistic regression (LR), decision tree (DT), linear support vector classifier (SVC), random forest (RF), support vector machine (SVM), gradient boosting regression tree, and multilayer perceptron models for data classification from keyboard strokes. The implemented model was 96.2% accurate when it came to stealing keyboard data. This means that cybercriminals can steal users' actual keyboard data in the real world with AI techniques. Trieu and Yang (2018) utilized Torch-rnn, an open-source machine learning technique to generate new candidate passwords based on a pattern similar to prior passwords. As

demonstrated in Figure 11, attackers can build new passwords by guessing one character at a time using the RNN, which is trained on previously obtained password sequences. The RNN produces a prediction by updating its hidden state at each timestamp by finding patterns over sequences. With his techniques, attackers are capable of constructing new terms that will present incredibly likely passwords. The comparisons were carried out for different dictionary lengths (i.e. the dictionary's total amount of words): 50, 100, 250, 500, 750, and 1000. Calculate the average of 100 trials for each trial. The result shows that the success rate of AI-driven password Brute-force attack outperformed the traditional algorithm Brute-force attack as shown in Figure 12. Figure 13 also illustrates the concept of AI-driven password brute-force attack (Trieu and Yang 2018), as illustrated in Figure 15.

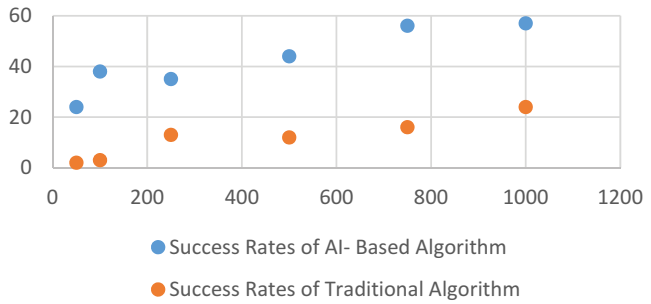

Figure 12. Success Rates of AI-Driven Password Brute-force Vs. Traditional Brute-force Attack.

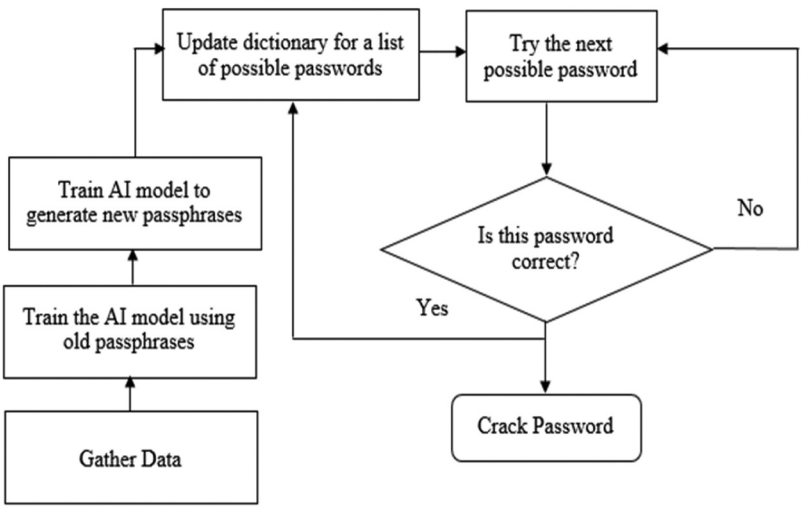

Figure 13. Password Brute-Force Attacks Powered by AI.

Intelligent Captcha Attack and Manipulation. Yu and Darling (2019) utilized an open-source Python Captcha package to boost recognition accuracy by combining TensorFlow object detection (TOD) and a speech segmentation method with CNN. The implemented model was able to determine which character was contained in a segmented sample. The result shows that the well-designed TOD+CNN model can crack open-source CAPTCHA libraries like Python Captcha and external captcha like the Delta40 benchmark. It has also been demonstrated that TOD+CNN can crack various types of CAPTCHAs, such as HashKiller13. Bursztein et al. (2014), developed a novel method for attacking captcha in a single step by combining segmentation and recognition problems using machine learning techniques. When both actions are done at the same time, the technique can take advantage of knowledge and context that would not be available if they were done separately. At the same time, it removes the need for any hand-crafted components, allowing this method to be applied to new Captcha schemes that the previous method could not. Without making any modifications to the algorithm or its settings, the authors were able to solve all of the real-world Captcha schemes they investigated exactly enough to consider the scheme insecure in reality, including Yahoo (5.33%) and ReCaptcha (33.34%). The success of this strategy against the Baidu (38.68%) and CNN (51.09%) schemes, both of which use occluding lines and character collapsing, implies that it can beat occluding lines in a broad sense. Noury and Rezaei (2020) proposed a vulnerability assessment Captcha solution based on deep learning. To explore the weaknesses and vulnerabilities of existing Captcha generation systems, the authors used a CNN model called DeepCAPTCHA. The numerical and alpha-numerical test datasets have cracking accuracy rates of 0.9894 and 0.983, respectively. That means more effort will be required to develop powerful Captchas that are resistant to AI-driven Captcha attack models.

Chen et al. (2018) suggested a hollow captcha attack that uses exact filling and nonredundant merging to improve attack accuracy and reduce attack time. To begin, the character shapes were methodically fixed using a thinning approach. Secondly, an inner-outer contour filling technique was developed for obtaining solid characters, which only fills the vacant character components rather than noise blocks. Finally, segmenting solid characters yields many distinct characters but only a few character components. Fourth, to obtain individual characters without duplication, a minimum-nearest neighbor merging technique was proposed. Finally, to obtain the final recognition results, (CNN) was used.

On text-based Captchas, Li et al.((2021) utilized cycle-GAN to train Captcha synthesizers to make several fake samples. Basic recognizers based on convolutional recurrent neural networks were trained using the fake dataset. After that, an active transfer learning mechanism optimizes the basic recognizer using small numbers of labeled real-world Captcha samples. This

method successfully solved the Captcha techniques used by ten (10) popular websites such as Amazon (88.4% success rate), Apple (0.877 success rate), Sina (0.85 success rate), Baidu (0.807 success rate), Weibo (0.798 success rate), eBay (0.743 success rate), Sogou (0.717 success rate), and Microsoft's two-layer approach had a success rate of 0.224, indicating that the attack is likely widespread. The findings demonstrate that combining multiple anti-recognition measures can increase captcha security, but only to a limited extent.

Gao et al. (2017) utilized deep learning to crack text-based Captchas and create image-based Captchas. As recognition engines, the authors used four CNN models with 2, 3, 5, and 6 convolutional layers. With success rates ranging from 0.10 to 0.90, the authors were able to defeat the Roman character-based Captchas used by the fifty most popular websites in the world, as well as three Chinese Captchas that use a broader character set. The average pace of this attack is substantially faster than prior attacks. Because these focused tactics cover almost all known resistance mechanisms, this offensive AI technique can breach other existing Captchas. Ye et al. (2018) presented a GAN-based approach for text-based Captcha solver. This was accomplished by first employing a Captcha synthesizer that generates synthetic captchas automatically so that a base solver may be learned, and then using transfer learning to fine-tune the basic solver on a limited set of real Captchas. The authors evaluated the implemented model on 33 different Captcha schemes, including 11 that are presently employed by 32 of the top 50 most visited websites. The authors demonstrated that their method is incredibly effective, cracking state-of-the-art Captchas in under 0.05 seconds. Gao et al. (2017) proposed a simple yet effective AI-driven solution for defeating Microsoft's two-layer captcha. The authors created an improved LeNet-5, a radical CNN model, as the recognition engine. The implemented model had a success rate of 44.6% and an average speed of 9.05 seconds on a standard desktop computer with a 3.3 GHz Intel Core i3 CPU.

Smart Fake Review Generation. Yao et al. (2017) presented a two-phase review generation and customization attack that can generate reviews that are unrecognizable by statistical detectors. The authors implemented an RNN-based fake review generation that is capable of generating misleading but realistic-looking reviews aimed at restaurants on the Yelp App. The result showed that the difficulty in detecting this form of attack is evidenced by the high quality of reviews generated.

AI-Model Manipulation. Malicious actors can purposely manipulate the data of machine learning models with adversary techniques to undermine the model. In several instances, malicious actors can insert a fake input set or manipulate the text of spam e-mails to bypass the spam filters, classification model. Cybercriminals can utilize a Naïve Bayes (NB) model that is used for

spam mail filtering by altering the input and training data to bypass the spam filter (Dheap 2017; Truong et al. 2020). Zhou et al. (2021) conducted a thorough investigation on deep model poisoning attacks on federated learning. The authors utilized the regularization term in the objective function to inject malicious neurons in the redundant network space to improve poisoning attacks in persistence, effectiveness, and robustness. DNNs models are subject to purposefully manipulated samples known as adversarial instances. These adversarial examples are created with little changes, yet they can cause DNN models to make incorrect predictions.

AI-Driven Attacks in the Delivery Stage of the Cybersecurity Kill Chain

From the selected studies, AI-driven concealment and AI-driven evasive attacks were identified as discussed below.

Intelligent Concealment and Evasive Malware. Bahnsen et al. (2018) utilized LSTM to generate sophisticated phishing URLs that are sufficient enough to be undetected by state-of-the-art cybersecurity detection infrastructures. Hu and Tan (2021) proposed a GAN technique that is capable of generating undetectable adversarial malware to bypass machine learning black-box detection systems. Anderson, Woodbridge, and Filar (2016) proposed a GAN-based automatic generation of undetectable malware URL that learns to bypass DNNs-based detection systems. The result shows that domains generated from the implemented GAN model bypass the DNNs, and GAN malware detection systems. Also, a random forest classifier that relies on hand-crafted features was easily bypassed. Kirat, Jang, and Stoecklin (2018) proposed a sophisticated evasive malware that is capable of hiding its malicious payload attack in video conferencing applications without being detected. The authors utilized DNNs to conceal its nefarious aim and only enable them for selected targets.

AI-Driven Attacks in the Exploitation Stage of the Cybersecurity Kill Chain

The selected studies identified AI-driven exploitation attacks, also known as AI-automated malware, as discussed below.

Behavioral Analysis to Find New Ways to Exploit. Petro and Morris (2017) evolved a Machine Learning model called DeepHack. The authors demonstrated how the implemented model could be utilized to break and bypass web-based applications using NNs and reinforcement learning (RL). Chung, Kalbarczyk, and Iyer (2019) demonstrated how k-means clustering could be utilized to determine attack effects on the target system, the author's utilized logical control data from the targeted system and a Gaussian distribution. The idea behind this technique was that once malware had gained access to a system, it needed to know how to operate and exploit weaknesses without the

need for further assistance from the attacker. The model effectively illustrated how AI-driven malware might launch their own attacks using a self-learning algorithm, reducing the level of knowledge needed by the attacker to successfully influence and exploit the target system.

Automated Disinformation Generation. Seymour and Tully (2016) demonstrated the automation of malicious payload in the phishing process by utilizing data science techniques to the target audience with personalized phishing messages. As a result, this malicious technique learned how to post phishing messages aimed just at high-value users, resulting in an automated targeted spear-phishing campaign. To find the high-value targets, the authors used the k-means clustering technique to cluster a collection of Twitter accounts into groups based on public profiles and social engagement indicators such as retweets, likes, and a number of followers. The attack disseminates customized, computer-generated posts with a truncated URL inserted in them once targets have been identified and established. NLP was used to determine which topics the target is interested in. As a result, it uses both Markov models and LSTMs to construct the content of the postings and also learns to guess the next word by analyzing the preceding context in the target's posting history.

AI-Driven Attacks in the Command and Control Stage of the Cybersecurity Kill Chain

Malicious actors commonly try to establish channels for the further communications link between it and the target with the objective of exerting influence over the compromised computer infrastructure and other systems on its internal network infrastructures. However, by utilizing AI techniques, cybercriminals do not require a C2 channel to execute their attacks (Kirat, Jang, and Stoecklin 2018). Based on the existing target attributes, the AI-Driven C2 malware automatically predicts when it will be unlocked across various sorts of nodes. As a result, a multi-layered AI-driven attack is capable of providing access to other computer infrastructure components remotely and automatically. Depending on the intents of the attacker, a successful AI-driven C2 attack can be used to disseminate the virus to other computers on the network, prompting the target to establish botnets, and downloading and installing remote access trojans (Zouave et al. 2020).

Intelligent Self-Learning Malware. Self-learning malware could be used to infiltrate cybersecurity defense systems in a supercomputer facility indirectly by interfering with the cyber-physical systems (CPS) automation system of the building (Chung, Kalbarczyk, and Iyer 2019). To classify the target system's logical control data and determine attack effects on the target system, the authors employed k-means clustering and Gaussian distribution. The goal of the simulation was to teach malware how to behave without the help of the

attacker once it had gained access to the system. The self-learning malware can use a self-learning algorithm to carry out its own attack plans, reducing the amount of information necessary by malicious actors to successfully manipulate the target system.

Automated Domain Generation. DGA classifiers employ GAN to evaluate and grade DNS queries executed by compromised hosts that were successful based on the values generated from the different training sets (Anderson, Woodbridge, and Filar 2016). DGAs are identified as queries that fall below a specific threshold and are prohibited. Anderson, Woodbridge, and Filar (2016) demonstrated how Cybercriminals could utilize domain Generation Algorithms (DGAs) to carry out sophisticated cyberattacks in the C2 phase of the cybersecurity kill chain. Malicious actors can also utilize this technique to establish data exfiltration (Sood, Zeadally, and Bansal 2017).

AI-Driven Attacks in the Action on Objective Stage of the Cybersecurity Kill Chain

AI-Driven DDoS Attack. There were automatic malware distribution and vulnerability type changes via C&C servers, but there was still a limitation. The requirement for human intervention was this constraint. The rise of AI in DDoS ushers is a new era of attack that does not necessitate the presence of humans. AI-Driven DDoS attack eliminates the need for human intervention entirely. Machines are now assaulting applications and state-of-the-art cybersecurity defense infrastructures. They are completely automated, altering vulnerability types and attack vectors in response to the defense's response. If one attacking signature fails, the machine can think for itself and switch to a different signature. All of this is carried out automatically, without the need for human intervention (Kaloudi and Li 2020; Kirat, Jang, and Stoecklin 2018).

RQ2: Traditional Targeted Cyberattacks and AI-Driven Cyberattacks

The traditional targeted cyberattack is a simplistic if-then conditional construct where it asks this question; is this a target? And if the answer is "No" the malicious program is going to end, and if the answer is "Yes" the malicious program is going to execute its attack (Kirat, Jang, and Stoecklin 2018). Figure 14 illustrates the decision logic of the traditional targeted attack.

Since cybercriminals realized that cybersecurity experts are using sandboxes to analyze and combat these traditional targeted attacks, they are now transforming this simplistic form of the if-then conditional construct to a very convoluted and complicated decision logic using Deep Neural Networks (DNN). With the concept of DNN, malicious actors can decide whether to attack or not. The problem for the defender is that it will be extremely difficult

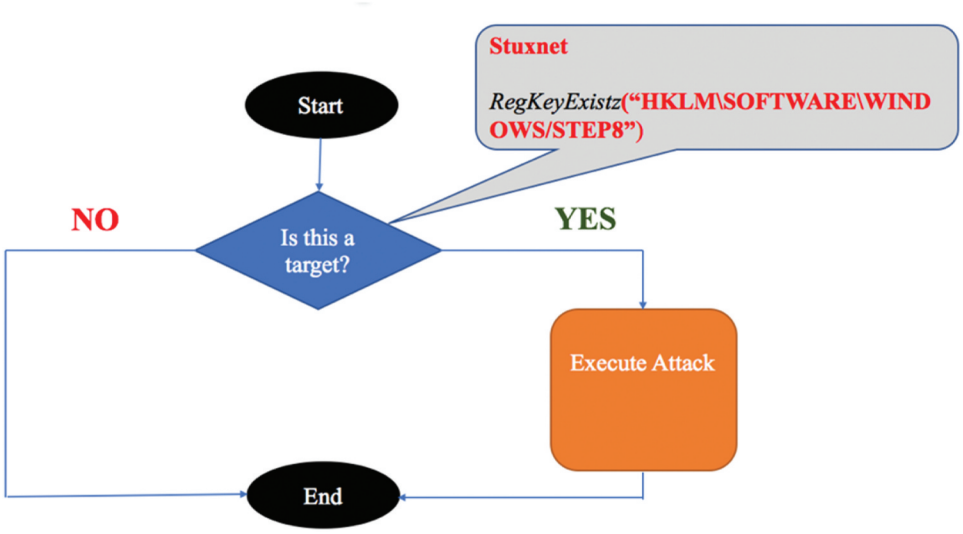

Figure 14. Traditional Targeted Cyberattack Decision Logic.

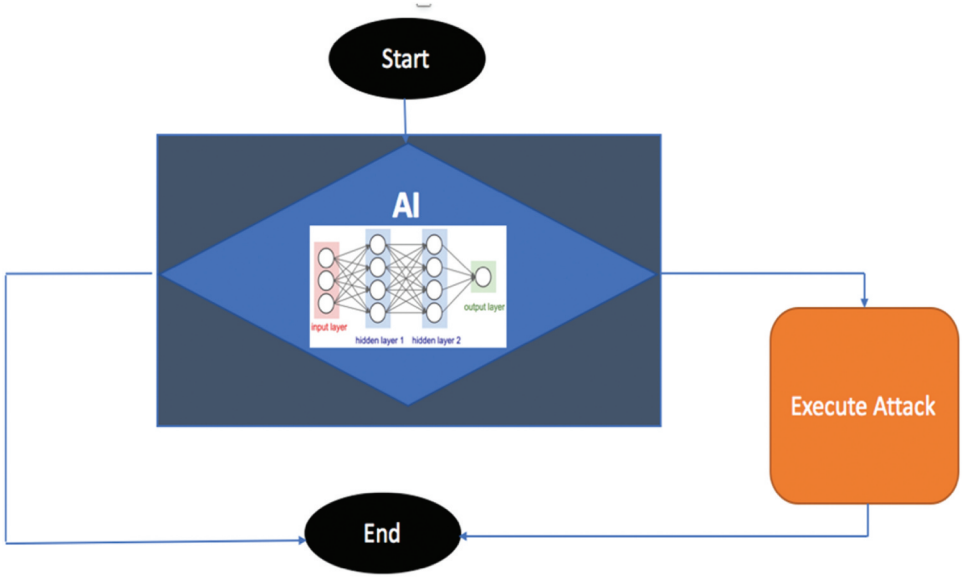

Figure 15. DNN Cyberattack Decision Logic.

to figure out what is the actual malicious code and what it is the right target (Kirat, Jang, and Stoecklin 2018). Figure 15 illustrates a DNN cyberattack decision logic.

RQ3: Impact of AI-Driven Cyberattacks

The consequences of these emerging AI-driven cyberattacks could be life-threatening and highly destructive. By undermining data confidentiality and integrity, highly sophisticated and stealthy attacks will erode trust in organizations and perhaps result in systemic failures. Consider a medical expert or physician giving a diagnosis based on tampered medical data or an oil rig drilling for crude in the wrong spot based on inaccurate geo-prospecting information. The potential of AI to learn and adapt has ushered in a new era of scalable, highly targeted, and human-like attacks. A smart and hostile offensive AI-driven attack will be able to adapt as it learns from its surroundings, allowing it to easily infect systems with little possibility of detection (Dixon and Eagan 2019). An AI-driven attack such as PassGAN is capable of generating a large number of efficient password guesses bypassing existing cybersecurity authentication infrastructures and causing greater damages without being noticed (John and Philip 2018).

Discussion

As discussed in this study, malicious actors are beginning to utilize AI-advanced data mining capabilities to execute more informed decisions. Learning from contextual data will specifically mimic trusted features of cyberspace or target weak points it discovers. This will enable AI-driven cyberattacks to avoid detection and maximize the damage they inflict on cyberspace. AI-driven attacks will be able to evolve as they learn from their surroundings, allowing them to effortlessly compromise systems with little possibility of detection. In general, it's apparent that AI-driven cyberattacks will only worsen, then it will be almost impossible for traditional cybersecurity tools to detect them It's simply a question of machine efficiency versus human labor. AI-driven threats will harness a multitude of cyberspace and computer resources well beyond what a human could enlist, resulting in an attack that is faster, more unpredictable, more sophisticated than even the strongest cybersecurity team can respond against. However, by using AI to fight AI, cybersecurity researchers, organizations, cybersecurity experts, and Government institutions can begin to prepare more advanced and sophisticated countermeasures to combat AI-driven attacks. The best method to prepare for this right now is to harden cybersecurity defense infrastructures to the best of their capacity, with the lowest possible number of false positives and negatives.

Conclusion

Cybercriminals are constantly changing and improving their attack efficiency, emphasizing the use of AI-driven techniques in the attack process. This study investigates the offensive capabilities of AI, allowing attackers to initiate attacks on a larger scale, with a broader scope, and at a faster pace. This

study reviewed existing literature on AI-driven cyberattacks, the improper use of AI in cyberspace, and the negative impact of AI-driven cyberattacks. The findings show that 56% of the AI-Driven cyberattack techniques identified were demonstrated in the access and penetration stage of the modified cybersecurity kill chain, 12% in the exploitation and C2 stage, 11% in the reconnaissance, and 9% in the delivery stage. CNN has the most appearances (five) among the AI techniques used by the selected authors to demonstrate access and penetration attacks. This study determined the status of existing AI-driven cyberattack research because 63% of current studies were based on implementation and evaluation, 25% on the proposed framework, and 12% on implementing AI techniques to execute AI-driven attacks. The findings show that traditional cybersecurity techniques' inability to detect and mitigate AI-driven attacks is directly related to their inability to cope with the speed, complex decision logic, and multiple variant nature of AI-driven attacks. With the emergence of these sophisticated attacks, organizations and security teams must quickly reform their strategies, be prepared to defend their digital assets with AI, and regain the advantage over this new wave of sophisticated attacks.

Finally, this study recommends that it is essential for the security research community, government, and cybersecurity experts to prepare and invest in advanced and sophisticated countermeasures to combat AI-driven cyberattacks and utilize AI to fight offensive AI. A trustworthy AI framework will be developed in the future to combat AI-Driven attacks while explaining essential features that influence the detection logic.

Disclosure Statement

No potential conflict of interest was reported by the author(s)

ORCID

Sanjay Misra http://orcid.org/0000-0002-3556-9331
Luis Fernandez-Sanz http://orcid.org/0000-0003-0778-0073
Vera Pospelova http://orcid.org/0000-0001-5801-1923

References

DarkTrace. 2021. The Next Paradigm Shift AI-Driven Cyber-Attacks. DarkTrace Research White Paper. https://www.oixio.ee/sites/default/files/the_next_paradigm_shift_-_ai_driven_cyber_attacks.pdf (accessed June 9, 2021).

Anderson, H. S., J. Woodbridge, and B. Filar. 2016. Deepdga: Adversarially-tuned domain generation and detection. In Proceedings of the ACM Workshop on Artificial Intelligence and Security, Vienna, Austria, 13–21.

Babuta, A., M. Oswald, and A. Janjeva. 2020. Artificial Intelligence and UK National Security Policy Considerations. Royal United Services Institute Occasional Paper.

Bahnsen, A. C., I. Torroledo, L. Camacho, and S. Villegas. 2018. DeepPhish: Simulating malicious AI. In APWG Symposium on Electronic Crime Research, London, United Kingdom, 1–8.

Bilal, M., A. Gani, M. Lali, M. Marjani, and N. Malik. 2019. Social profiling: A review, taxonomy, and challenges. *Cyberpsychology, Behavior and Social Networking* 22 (7):433–50. doi:10.1089/cyber.2018.0670.

Bocetta, S. 2020. Has an AI cyberattack happened yet? https://www.infoq.com/articles/ai-cyberattacks/ (accessed December 9,2020).

Brundage, M., S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfinkel, A. Dafoe, P. Scharre, T. Zeitzoff, B. Filar, et al. 2018. *The malicious use of artificial intelligence: forecasting, prevention, and mitigation*. Oxford: Future of Humanity Institute.

Bursztein, E., J. Aigrain, A. Moscicki, and J. C. Mitchell. 2014. The end is nigh: generic solving of text-based CAPTCHAs. 8th Usenix workshop on Offensive Technologies WOOT '14, San Diego, CA, USA.

Cabaj, K., Z. Kotulski, B. Księżopolski, and W. Mazurczyk. 2018. Cybersecurity: trends, issues, and challenges. *EURASIP Journal On Information Security*. doi:10.1186/s13635-018-0080-0.

Cani, A., M. Gaudesi, E. Sanchez, G. Squillero, and A. Tonda (2014). Towards automated malware creation. Proceedings of The 29Th Annual ACM Symposium On Applied Computing, Gyeongju Republic of Korea, 157–60. doi: 10.1145/2554850.2555157.

Chakkaravarthy, S. S, D. Sangeetha, V. M. Rathnam, K. Srinithi, and V. Vaidehi. 2018. Futuristic cyber-attacks. *International Journal of Knowledge-Based and Intelligent Engineering Systems* 22 (3):195–204. doi: 10.3233/kes-180384.

Chen, J., X. Luo, J. Hu, D. Ye, and D. Gong. 2018. An Attack on Hollow CAPTCHA Using Accurate Filling and Nonredundant Merging. *IETE Technical Review*, 35(sup1):106–118. doi:10.1080/02564602.2018.1520152.

Chung, K., Z. T. Kalbarczyk, and R. K. Iyer. 2019. Availability attacks on computing systems through alteration of environmental control: Smart malware approach. Proceedings of the 10th ACM/IEEE International Conference on Cyber-Physical Systems, Montreal Quebec, Canada, 1–12.

Dheap, V. 2017. AI in cybersecurity: A balancing force or a disruptor? https://www.rsaconference.com/industry-topics/presentation/ai-in-cybersecurity-a-balancing-force-or-a-disruptor (accessed February 13, 2020).

Dixon, W., and N. Eagan. 2019. AI will power a new set of tools and threats for the cybercriminals of the future. https://www.weforum.org/agenda/2019/06/ai-is-powering-a-new-generation-of-cyberattack-its-also-our-best-defence/ (accessed December 3, 2020).

Fischer, E. 2016. Cybersecurity issues and challenges: In brief. . https://fas.org/sgp/crs/misc/R43831.pdf (accessed December 1, 2020).

Gao, H., M. Tang, Y. Liu, P. Zhang, and X. Liu. 2017. Research on the security of microsoft's two-layer captcha. *IEEE Transactions On Information Forensics And Security* 12 (7):1671–85. doi:10.1109/tifs.2017.2682704.

Hamadah, S., and D. Aqel. 2020. Cybersecurity becomes smart using artificial intelligent and machine learning approaches: An overview. *ICIC Express Letters, Part B: Applications* 11 (12):1115–1123. doi:10.24507/icicelb.11.12.1115.

Hitaj, B., P. Gasti, G. Ateniese, and F. Perez-Cruz. 2019. PassGAN: A deep learning approach for password guessing. *Applied Cryptography and Network Security* 11464:217–37. doi:10.1007/978-3-030-21568-2_11.

Hu, W., and Y. Tan. 2021. Generating adversarial malware examples for black-box attacks based on GAN.https://arxiv.org/abs/1702.05983 (accessed August 12, 2021).

John, S., and T. Philip. 2018. Generative models for spear phishing posts on social media. NIPS Workshop On Machine Deception, California, USA. arXiv:1802.05196

Kaloudi, N., and J. Li. 2020. The AI-based cyber threat landscape. *ACM Computing Surveys* 53 (1):1–34. doi:10.1145/3372823.

Kirat, D., J. Jang, and M. Stoecklin. 2018. DeepLocker concealing targeted attacks with AI locksmithing. https://www.blackhat.com/us-18/briefings/schedule/index.html#deeplocker—concealing-targeted-attacks-with-ailocksmithing-11549 (accessed December 4, 2020).

Lee, K., and K. Yim. 2020. Cybersecurity threats based on machine learning-based offensive technique for password authentication. *Applied Sciences* 10 (4):1286. doi:10.3390/app10041286.

Li, C., X. Chen, H. Wang, P. Wang, Y. Zhang, and W. Wang. 2021. End-to-end attack on text-based CAPTCHAs based on cycle-consistent generative adversarial network. *Neurocomputing* 433:223–36. doi:10.1016/j.neucom.2020.11.057.

Meng, G., Y. Xue, C. Mahinthan, A. Narayanan, Y. Liu, J. Zhang, and T. Chen. 2016. Mystique. Proceedings of the 11Th ACM On Asia Conference On Computer and Communications Security, Xi'an, China, 365–76. doi:10.1145/2897845.2897856.

Moher, D., A.Liberati, J.Tetzlaff, and D. G.Altman. 2010. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. International Journal of Surgery 8(5): 336–341. doi:10.1016/j.ijsu.2010.02.007.

Ney, P., K. Koscher, L. Organick, L. Ceze, and T. Kohno. 2017. Computer security, privacy, and dna sequencing: compromising computers with synthesized DNA, privacy leaks, and more. USENIX Security Symposium, Vancouver, BC, Canada, 765–79.

Noury, Z., and M. Rezaei. 2020. Deep-CAPTCHA: A deep learning based CAPTCHA solver for vulnerability assessment. ArXiv, abs/2006.08296.

Petro, D., and B. Morris. 2017. Weaponizing machine learning: Humanity was overrated anyway. DEF CON.

Rigaki, M., and S. Garcia. 2018. Bringing a GAN to a knife-fight: adapting malware communication to avoid detection. IEEE Security and Privacy Workshops (SPW), San Francisco, CA, USA. doi:10.1109/spw.2018.00019.

Seymour, J., and P. Tully. 2016. Weaponizing data science for social engineering: Automated E2E spear phishing on Twitter. https://www.blackhat.com/docs/us-16/materials/us-16-Seymour-Tully-Weaponizing-Data-Science-For-Social-EngineeringAutomated-E2E-Spear-Phishing-On-Twitter-wp.pdf (accessed December 21, 2020).

Sood, A., S. Zeadally, and R. Bansal. 2017. Cybercrime at a Scale: A Practical Study of Deployments of HTTP-Based Botnet Command and Control Panels. *IEEE Communications Magazine* 55 (7):22–28. doi:10.1109/mcom.2017.1600969.

Tang, M., H. Gao, J. Yan, F. Cao, ZhangZ., L. Lei, M. Zhang, P. Zhou, X. Wang, X. Li, and L. X. Jiawei. 2016. A simple generic attack on text captchas. Proceedings 2016 Network And Distributed System Security Symposium, San Diego, California. doi:10.14722/ndss.2016.23154.

Thanh, C., and I. Zelinka. 2019. A survey on artificial intelligence in malware as next-generation threats. *MENDEL* 25 (2):27–34. doi:10.13164/mendel.2019.2.027.

Trieu, K., and Y. Yang. 2018. Artificial intelligence-based password brute force attacks. Proceedings of Midwest Association for Information Systems Conference, St. Louis, Missouri, USA, 13(39).

Truong, T., I. Zelinka, J. Plucar, M. Čandík, and V. Šulc. 2020. Artificial intelligence and cybersecurity: past, presence, and future. *Advances In Intelligent Systems And Computing* 351–63. doi:10.1007/978-981-15-0199-9_30.

Usman, M., M. Jan, X. He, and J. Chen. 2020. A survey on representation learning efforts in cybersecurity domain. *ACM Computing Surveys* 52 (6):1–28. doi:10.1145/3331174.

Xu, W., D. Evans, and Y. Qi. 2018. Feature squeezing: Detecting adversarial examples in deep neural networks. Proceedings 2018 Network and Distributed System Security Symposium, San Diego, California, USA. doi:10.14722/ndss.2018.23198.

Yao, Y., B. Viswanath, J. Cryan, H. Zheng, and B. Zhao. 2017. Automated crowdturfing attacks and defenses in online review systems. Proceedings Of The 2017 ACM SIGSAC Conference On Computer And Communications Security, Dallas Texas, USA. doi:10.1145/3133956.3133990.

Ye, G., Z. Tang, D. Fang, Z. Zhu, Y. Feng, P. Xu, X. Chen, and Z. Wang. 2018. Yet another text captcha solver. Proceedings of The 2018 ACM SIGSAC Conference On Computer And Communications Security, Toronto, Canada. doi:10.1145/3243734.3243754.

Yu, N., and K. Darling. 2019. A low-cost approach to crack python CAPTCHAs using AI-based chosen-plaintext attack. *Applied Sciences* 9 (10):2010. doi:10.3390/app9102010.

Zhou, X., M. Xu, Y. Wu, and N. Zheng. 2021. Deep model poisoning attack on federated learning. *Future Internet* 13 (3):73. doi:10.3390/fi13030073.

Zouave, E., M. Bruce, K. Colde, M. Jaitnee, I. Rodhe, and T. Gustafsson. 2020. Artificially intelligent cyberattacks.https://www.statsvet.uu.se/digitalAssets/769/c_769530-l_3-k_rapport-foi-vt20.pdf (accessed December, 21 2020).