Artificial Intelligence Powered Cyber Attacks

Aakanksha[1], Sanjana Dwivedi[1], Sonia[1], Ravish Sharma[2*]

ABSTRACT

Artificial intelligence (AI) has emerged as a key component of modern cybersecurity, providing advanced threat detection and system defense solutions. However, enemies are using the same technology that enables defenders to launch complex cyberattacks. This paper examines the dual role of AI in cybersecurity, studying its applications in defense mechanisms and its misuse in malicious activities, and also highlights how urgently ethical AI research, continuous advancement in AI-driven security measures, and cooperative efforts are needed to prevent the growing risks in the digital world. Key issues highlighted here include AI-driven threat detection, the growth of AI-powered cyberattacks, the challenges faced by deepfake technologies, adversarial AI strategies, and the developments of AI-enhanced ransomware.

Keywords : *Artificial Intelligence, Cyber Security, Cyber Attacks*

1. Introduction

AI has advanced from simple rule-based systems to complex models that learn, adapt, and make independent decisions. AI is now an integral part of modern life, influencing areas such as personal assistants (e.g., Alexa and Siri), automated customer service, healthcare, and cybersecurity. While AI enhances security frameworks, it also presents significant threats when used by cybercriminals.

The different AI systems known to humanity, categorized into four types based on their functionalities are discussed below[1]:

Reactive machines: These machines cannot learn from the previous tasks and cannot function based on memory. Reactive machines are the basic type of AI systems that do not have memories and cannot save past experiences, which in turn can be used to make future decisions. We cannot expect these reactive machines to interact, and show emotions or consciousness but are reliable in completing tasks as they respond in the same way for the same situation every time. Reactive machines are made for specific purposes, hence is easy to trick them.

Limited Memory: Limited Memory Machines have short-term memory that enables them to store acquired experiences temporarily and take actions/decisions based on them. The data gathered from previous experiences is not stored in the machine's content library so that it can be used in subsequent situations. This type of AI is more advanced than Reactive machines and as their name suggests, they can look into the past and create a memory- but not the distant past.

Theory of Mind: The AI machines discussed above exist and are used in everyday life whereas Theory of Mind (ToM) is just a theory for now and it's the future we await eagerly. These types of machines will be programmed to function on the understanding of human thoughts and emotions that in turn influence their decision-making and behaviour. This concept is based on psychology and is known as the Theory of Mind.

Self-Aware: We define someone who is Self-Aware as who is aware of one's personality and individuality. Similarly, a Self-aware AI would be a system aware of itself and its internal states.

There are new cybersecurity dangers as AI systems get more complex. Skipping traditional security measures, AI-powered attacks can automate harmful actions, find weaknesses in systems, and carry out complex attacks that change over time. This paper discusses about AI's role in cybersecurity, highlighting mainly its capacity for harmful usage.

2. AI in Cybersecurity

There are two sides to artificial intelligence (AI) in the field of cybersecurity (Alzboon et al., 2023). It provides strong defences against online threats, but when bad actors use AI

[1] https://www.ibm.com/think/topics/artificial-intelligence-types

1. *Department of Computer Science, SRCASW, University of Delhi*
2. *Department of Computer Science, PGDAV College, University of Delhi*
* *Corresponding Author ✉ ravish.sharma@pgdav.du.ac.in*

Received: 10 February, 2025
Available online: 15 April, 2025

to carry out increasingly complex attacks, it also creates new difficulties. A scientific arms race has been created by this changing environment, in which attackers and defenders are constantly changing their tactics.

Cybercriminals are using AI to get beyond security measures that enterprises are increasingly depending on for threat detection and response. These days, sophisticated machine learning algorithms can be used to automate flaw recognition, produce more convincing phishing emails, and even impersonate human behavior to avoid detection. As a result of the application of AI to cybersecurity, adversarial machine learning techniques have emerged in which attackers try to trick or control AI systems. This involves creating inputs that take advantage of flaws in machine learning models or manipulating data to trick anomaly detection systems. AI-powered malware can also change its behavior according to its surroundings, which makes it harder for conventional security procedures to detect and eliminate threats. These clever, harmful programs are able to change their code, expected defensive moves, and learn from failed attempts.

2.1 AI as a defense mechanism in Cybersecurity

AI helps in automating repetitive tasks and manually intensive tasks, which results in freeing up time for security analysts to focus on more complex problems. AI is utilized in spam filtering, endpoint security, phishing detection, and various other applications. AI can analyse a large amount of data, identify different patterns, and adapt over time to improve its capabilities (Kavitha & Thejas, 2024)AI can increase threat detection by analyzing data such as typing styles, fingerprints, and voice patterns to authenticate users and identify potential threats. It helps in finding the characteristics of cyberattacks and also helps in strengthening defenses. (Morovat & Panda, 2020)

2.2 AI as a threat to Cybersecurity

The attacks performed or executed using AI algorithms and techniques are called **AI-powered Cyber attacks** that accelerate, automate, and enhance various phases of cybercrimes. They work by identifying vulnerabilities in a system, deploying and advancing attack paths, exfiltrating or tampering with data, and establishing backdoors within systems (Ilieva & Stoilova, 2024). Over time, these AI-powered cyberattacks can evolve to prevent detection or create a pattern that cannot be identified and detected by a security system. Conversely, cybercriminals exploit AI to conduct more sophisticated and automated attacks (Rahman et al., 2023)Cybercriminals use AI to automate and enhance various phases of their attacks (Morovat & Panda, 2020). AI algorithms can learn and evolve from them, which results in making these attacks more sophisticated, better, and harder to detect (Tsikerdekis et al., 2024). To get a better understanding of the topic 'AI-powered Cyberattacks', let us dive deep into its characteristics and understand how it works (A. Ali et al., 2023)

Reinforcement Learning: AI algorithms continuously learn and adapt to improve their techniques or avoid detection. This makes AI-powered cyberattacks more dynamic and challenging to counter (Oh et al., 2024).

Efficient Data Gathering: During the initial phase of a cyberattack, the attackers will search and gather data about their targets and their possible vulnerabilities and assets that can be compromised. AI can automate and accelerate this process, helping the attackers to shorten this research phase.

Employee Targeting and Customization: AI can collect and analyze information from public sources to create personalized messages for phishing attacks and other malicious activities. It can also identify high-value targets within an organization, making attacks more effective.

Automation and Speed: AI can automate various stages of a cyberattack, from surveillance to execution. This allows attackers to launch and manage large-scale attacks with minimal human interference, increasing the speed and productivity of their operations.

Adaptive Learning: AI algorithms can learn from previous attacks and adapt their approach in real time. This makes it harder for traditional security measures to detect and counter these attacks, as the AI can continuously evolve to ignore defenses.

Credential Stuffing and Brute Force Attacks: AI can strengthen credential stuffing and brute force attacks by quickly testing a large number of username and password combinations. Machine learning algorithms can identify patterns and guess likely password variations, increasing the chances of a successful breach.

Botnets and Distributed Attacks: AI can manage and control botnets more productively, coordinating distributed denial-of-service (DDoS) attacks with greater precision. This makes it possible to overpower targets with traffic, causing significant disruptions.

Zero-Day Exploits: AI can assist in exploring and exploiting zero-day vulnerabilities, which are previously unknown security flaws. By observing and analysing software and systems for weaknesses, AI can identify potential entry points for attacks before they are patched (S. Ali et al., 2022).

Ransomware: AI can strengthen ransomware attacks by finding critical files and systems to target, optimizing encryption methods, and computerizing the ransom demand process. This increases the chances of a successful attack and higher ransom payments. These capabilities make AI-powered cyber attacks an important threat to organizations and individuals alike. Cyber security professionals need to stay ahead of these improvements by implementing robust security measures and continuously updating their defenses (Yarali et al., 2023).

These cyberattacks are categorized into different types based on the algorithms, tools, and techniques they employ. In the below sections, we will be discussing the types of AI-powered Cyberattacks and their tools.

2.3 Types of AI-Powered Cyberattacks

AI-powered cyberattacks can be broadly categorized into two main types, namely, Social Engineering attacks and Phishing attacks, as shown in figure 1 below. Phishing attacks are further classified into Deep fakes, Adversarial AI, Malicious GPTs, and Ransomware attacks (Guembe et al., 2022). This section briefly discusses these categories of AI-powered cyberattacks.

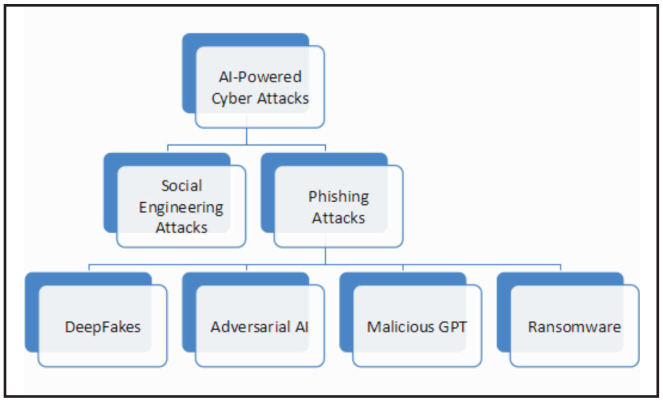

***Figure 1:** Types of AI-powered Cyber Attacks*[2]

Social engineering attacks

These types of attacks are used to assist in the research and execution of a social engineering attack that aims to manipulate a possible target to share sensitive information, transfer money or funds, or grant access to their system or device. The algorithm (Fakhouri et al., 2024) can be used to:

- Identify an ideal target and research its vulnerabilities
- Develop a fake but relevant persona to carry out the communications
- Develop realistic and believable scenarios to gain the attention of the attack target
- Create personalized messages or video or audio recordings to engage the target

Phishing attacks

AI-driven phishing attacks utilize Generative AI to craft fake personas and realistic communications, including emails, texts, videos, and phone calls. Their goal is to steal sensitive information like account and credit card details or to trick targets into installing malicious software.

The next few sections discuss **Deepfakes,** Adversarial **AI, Malicious GPT, and Ransomware attacks** (Schmitt & Flechais, 2024), the common Phishing attacks performed by the attackers, which use AI as the key tool.

3. Deepfake and AI-Generated Threats

One of the most alarming AI-driven threats is Deepfake technology, which utilizes advanced machine learning to create highly realistic but fake audio, video, and image content. Deepfake represent a captivating yet divisive application of artificial intelligence technology. This innovation employs deep learning algorithms to generate highly convincing but artificial videos, images, or audio recordings.

At the heart of the Deepfake creation process is data collection. A vast collection of datasets of images, videos, or audio clips are gathered to train the AI system. Then deep learning algorithms are employed to recognize patterns in facial expressions, voice modulation, and other physical attributes and train the system. Finally, deepfake content is generated by applying the learned data to new content, superimposing one individual's likeness onto another with remarkable realism.

The term "Deepfake," combining "deep learning" and "fake," denotes media that has been manipulated or created using AI (Yarali et al., 2023). Computer vision research has significantly focused on both the production and identification of Deepfakes (Oh et al.,2024). The creation of Deepfakes through deep learning techniques has led to their increasing realism and detection difficulty (Lyu, 2020). The background and state-of-the-art detection methods for Deepfakes have been extensively examined in multiple studies (Reiss et al., 2023).

Deepfake technology is used in various fields, including (Mahmud & Sharmin, 2021)

- **Entertainment:** Recreating actors and generating special effects in the film industry.
- **Misinformation:** Spreading fake news and manipulated media to influence public opinion and political outcomes.
- **Identity Theft & Fraud:** Impersonating individuals through AI-generated voices and videos, leading to financial fraud and reputational harm.
- **Cyber Espionage:** Enabling hackers to mimic high-profile individuals and gain unauthorized access to sensitive data.

Despite the rising threat of deepfake attacks, efforts are underway to combat misuse through advanced detection methods:

- **Deep Learning-Based Detection:** AI models analyze inconsistencies in facial expressions, voice modulation, and image artifacts.
- **Media-Modality Fusion:** Integrating multiple media

[2] https://github.com/cybershujin/Threat-Actors-use-of-Artifical-Intelligence/blob/main

forms – images, videos, and audio – to enhance detection accuracy.

This approach integrates various media forms, such as images, videos, and audio, to boost detection accuracy (Gupta et al., 2023).

Many problems still exist even after deepfake attack detection advances and constantly changing AI solutions. A major challenge in the identification of deepfakes is the ongoing development of generative algorithms. Deepfake producers update their methods as detection techniques advance, leading to constant technological competition. To remain effective in this dynamic environment, detection techniques must always be modified and innovative. The lack of large, diverse, and high-quality datasets for deepfake detection system training and evaluation is another challenge. Even though deepfake generation is capable of producing a wide variety of amazingly realistic content, compiling an extensive dataset that captures this range of diversity is still very difficult. Another challenge is ***evasion attacks***. Deepfakes can be engineered to bypass detection methods, necessitating the development of more robust identification techniques. Deepfake generation techniques can produce highly realistic and diverse content, but obtaining a comprehensive dataset that captures this variability is challenging. The potential misuse of deepfakes for various nefarious purposes, including political manipulation, financial deception, and non-consensual pornography, further complicates the issue. Researchers must grapple with the ethical considerations of their work, striving to develop solutions that strike a balance between safeguarding privacy, ensuring security, and preserving freedom of expression.

4. Adversarial AI and Model Manipulation

Adversarial AI, or adversarial machine learning, involves undermining the effectiveness of AI/ML systems through manipulation. This can happen at different stages, such as tampering with training data, poisoning ML models, or generating deceptive inputs that lead to incorrect outputs. Essentially, adversarial AI/ML exploits the decision-making process of an AI system to bypass security measures in a trained and operational model. (Schmitt & Flechais, 2023).

These attacks can be done in two ways: data poisoning and model tampering.

Data poisoning

As the training data is very critical part of the performance of the ML model, it is therefore an attractive target in Adversarial AI. Data or AI poisoning attacks are deliberate attempts to alter and manipulate the training data of AI and ML models to corrupt their behaviour and cause biased/undesirable outputs. A variety of methods can be used by the attackers to execute Data Poisoning which includes Mislabel attack (Mislabel portions of the AI model's training data set leading to incorrect patterns of Output), Data Injection (Injecting malicious data samples in the AI/ML training data models), Data manipulation (Altering the Data within the Training Data set) and, Backdoors (Planting a Hidden Vulnerability) (Rahman et al., 2023)

Model Tampering

Adversarial AI also uses tampering with ML models in which unauthorized modifications are made to an ML model's parameters or structure. It can take many forms, which are listed below:

- Manipulation of Data
- Deletion of Data
- Insertion of Data
- Duplication of Data
- Substitution of Data
- Replay attacks
- Spoofing

5. Malicious GPT

A transformer-based language model called the Generative Pre-trained Transformer is at the beginning of a new era of technological innovation brought about by the rapid developments in artificial intelligence. Often known as GPT, these models have proven to be particularly successful at a variety of activities, ranging from scientific work and music composition to text and image generation. But, as their potential for misuse has become more obvious, the strength of GPT models also comes with a price. Concerns over the possibility of harmful uses of these technologies have increased with the recent release of ChatGPT, a very advanced conversational AI helper (Gupta et al., 2023). Nguyen et al.'s study has shown the extremely complex phishing attempts, highlighting the necessity of improved defenses in these AI systems. Researchers have also shown that there are still ways for GPT models to produce offensive and dangerous content, even with ongoing efforts to create safe and ethical conversational AI. These results highlight the urgent need to address these complex and diverse issues raised by the widespread use of GPT models since their potential for misuse poses a serious risk to people, institutions, and society as a whole (Iqbal and Rahbi, 2025). Table 1 shows various types of Malicious GPTs.

6. AI-Driven Ransomware Attacks

Ransomware is a specific type of malware designed to prevent users or organizations from accessing files on their computers. As its name implies, cyber attackers encrypt these files using their algorithms and keys, demanding a ransom payment for the decryption key. The overwhelming nature of such attacks makes paying the ransom seem like the easiest and most cost-effective way to regain access to important files. In recent years,

Table 1 : *Types of Malicious GPT*[3]

WORMGPT	FRAUDGPT	POISONGPT	EVILGPT	XXX-GPT	WOLFGPT
It is without ethical filters. Any low-skill hacker can use it.	New & exclusive bot designed for fraudsters, hackers, spammers, and like-minded individuals. most advanced bot of its kind	Focus is spreading misinformation Online tool inserts false details regarding historical events	A more powerful alternative to WormGPT.	offers a various range of malicious services too: deploy botnets, RATs, malware, key loggers, infostealers, you name it	promises amazing evasion capabilities and the possibility to generate a variety of malicious content types

Table 2 : *Types of Ransomware Attacks*[4]

Lockers	Lockers completely block access to your system, rendering your files and applications unavailable. A lock screen shows the ransom demand, often with a countdown clock to create urgency and pressure victims to act quickly.
Scareware	Scareware is fraudulent software that falsely claims to find a virus or problem on your computer, prompting you to pay for a solution. Some scareware may lock your computer, while others generate numerous pop-up alerts without harming your files.
Crypto Ransomware	Crypto ransomware attacks involve demanding ransom payments in cryptocurrency because these digital currencies are harder to trace and are not controlled by traditional financial systems.
Wiper	Wipers are a type of malware distinct from ransomware. Unlike ransomware, which holds files for ransom, wipers aim to permanently erase access to files, including deleting the only copy of the encrypted data.
Ransomware as a Service (RaaS)	RaaS is a malware distribution model where criminal groups provide affiliates to attackers. They collaborate to infect targets and split the ransom payments.
Doxware or Leakware	Leakware threatens to publish sensitive personal or company information online, causing many to panic and pay the ransom to protect their data. One type of this ransomware impersonates law enforcement, claiming illegal online activity was detected and offering to avoid jail time by paying a fine.

ransomware has rapidly emerged as the most prominent and visible form of malware, significantly impacting various sectors, including healthcare, public services, and numerous organizations worldwide.

A notable incident occurred on November 23, 2022, when the All India Institute of Medical Sciences (AIIMS) fell victim to a ransomware attack. This attack incapacitated computer-operated services at the government-run hospital for more than 15 days, forcing the hospital to revert to manual operations.

A summary of the most common types of ransomware attacks and how they work is shown in Table 2 below.

7. Conclusion

AI has transformed crimes and cybersecurity in both offensive and defensive ways. Cybercriminals use the same technology to carry out complex crimes, even though AI-powered security solutions strengthen defenses. The increasing complexity of cyber threats are shown by the rise of deepfakes, opposing AI, and AI-driven ransomware. In order to overcome these obstacles, enterprises, governments, and cybersecurity specialists must work together and continuously evolve AI-driven security measures and ethical AI development. Security experts can reduce risks and take advantage of AI's potential for a safer digital future by staying ahead of AI-powered dangers.

8. References

1. Ali, A., Khan, M. A., Farid, K., Akbar, S. S., Ilyas, A., Ghazal, T. M., & Al Hamadi, H. (2023). The Effect of Artificial Intelligence on Cybersecurity. 2nd International Conference on Business Analytics for Technology and Security, ICBATS 2023. https://doi. org/ 10.1109 /ICBATS 57792.2023.10111151

[3] https://github.com/cybershujin/Threat-Actors-use-of-Artifical-Intelligence/blob/main /Dark%20LLMs%20and%20Malicious%20AIs.MD

[4] https://www.crowdstrike.com/en-us/cybersecurity-101/ransomware/types-of-ransomware/

2. Ali, S., Rehman, S. U., Imran, A., Adeem, G., Iqbal, Z., & Kim, K. Il. (2022). Comparative Evaluation of AI-Based Techniques for Zero-Day Attacks Detection. Electronics 2022, Vol. 11, Page 3934, 11(23), 3934. https://doi.org/10.3390/Electronics11233934
3. Alzboon, M. S., Bader, A. F., Abuashour, A., Alqaraleh, M. K., Zaqaibeh, B., & Al-Batah, M. (2023). The Two Sides of AI in Cybersecurity: Opportunities and Challenges. Proceedings of 2023 2nd International Conference on Intelligent Computing and Next Generation Networks, ICNGN 2023. https://doi.org/10.1109/ICNGN 59831.2023.10396670
4. Fakhouri, H. N., Alhadidi, B., Omar, K., Makhadmeh, S. N., Hamad, F., & Halalsheh, N. Z. (2024). AI-Driven Solutions for Social Engineering Attacks: Detection, Prevention, and Response. 2nd International Conference on Cyber Resilience, ICCR 2024. https://doi.org/10.1109/ICCR61006.2024.10533010
5. Guembe, B., Azeta, A., Misra, S., Osamor, V. C., Fernandez-Sanz, L., & Pospelova, V. (2022). The Emerging Threat of Ai-driven Cyber Attacks: A Review. Applied Artificial Intelligence, 36(1). https://doi.org/10.1080/08839514.2022.2037254
6. Gupta, M., Akiri, C., Aryal, K., Parker, E., & Praharaj, L. (2023). From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy. IEEE Access, 11, 80218–80245. https://doi.org/10.1109/ACCESS.2023.3300381
7. Ilieva, R., & Stoilova, G. (2024). Challenges of AI-Driven Cybersecurity. 2024 33rd International Scientific Conference Electronics, ET 2024 – Proceedings. https://doi.org/10.1109/ET63133.2024.10721572
8. Iqbal, J., & Rahbi, F. (2025). The Evolution of Ransomware: AI-Powered Detection and Prevention Strategies.
9. Kavitha, D., & Thejas, S. (2024). AI Enabled Threat Detection: Leveraging Artificial Intelligence for Advanced Security and Cyber Threat Mitigation. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3493957
10. Lyu, S. (2020). Deepfake detection: Current challenges and next steps. 2020 IEEE International Conference on Multimedia and Expo Workshops, ICMEW 2020. https://doi.org/10.1109/ICMEW46912.2020.9105991
11. Mahmud, B. U., & Sharmin, A. (2021). Deep insights of deepfake technology: A review. arXiv preprint arXiv:2105.00192.
12. Morovat, K., & Panda, B. (2020). A Survey of Artificial Intelligence in Cybersecurity. Proceedings – 2020 International Conference on Computational Science and Computational Intelligence, CSCI 2020, 109-115. https://doi.org/10.1109/CSCI51800.2020.00026
13. Oh, S. H., Kim, J., Nah, J. H., & Park, J. (2024). Employing Deep Reinforcement Learning to Cyber-Attack Simulation for Enhancing Cybersecurity. Electronics 2024, Vol. 13, Page 555, 13(3), 555. https://doi.org/10.3390/ Electronics13030555
14. Rahman, M. M., Siddika Arshi, A., Hasan, M. M., Farzana Mishu, S., Shahriar, H., & Wu, F. (2023). Security Risk and Attacks in AI: A Survey of Security and Privacy. Proceedings – International Computer Software and Applications Conference, 2023-June, 1834–1839. https://doi.org/10.1109/COMPSAC57700.2023.00284
15. Reiss, T., Cavia, B., & Hoshen, Y. (2023). Detecting Deepfakes Without Seeing Any. https://arxiv.org/abs/2311.01458v1
16. Schmitt, M., & Flechais, I. (2023). Digital Deception: Generative Artificial Intelligence in Social Engineering and Phishing. SSRN Electronic Journal. https://doi.org/10.2139/SSRN.4602790
17. Schmitt, M., & Flechais, I. (2024). Digital deception: generative artificial intelligence in social engineering and phishing. Artificial Intelligence Review 2024 57:12, 57(12), 1-23. https://doi.org/10.1007/S10462-024-10973-2
18. Tsikerdekis, M., Zeadally, S., & Katib, I. (2024). Defenses Against Artificial Intelligence Attacks. Computer, 57(11), 49–59. https://doi.org/10.1109/MC.2024.3420782
19. Yarali, A., Rodocker, E., & Gora, C. (2023). Artificial Intelligence in Cybersecurity: A Dual-Nature Technology. 2023 International Conference on Computational Science and Computational Intelligence (CSCI), 234-240. https://doi.org/10.1109/CSCI62032.2023.00042

❖ ❖ ❖