ISSN: 3049-1479(Online)

A Comprehensive Study of Cyber Attacks against Artificial Intelligence Systems

Aayush Shah[1], Momin Samir[2], Deep Solanki[3], JigarGajjar[4]

Lok Jagruti Kendra University, Ahmedabad, India[1]
Lok Jagruti Kendra University, Ahmedabad, India[2]
TechD Cybersecurity Limited, Ahmedabad, India[3]
TechD Cybersecurity Limited, Ahmedabad, India[4]

a00311027@gmail.com[1], mominsamir2202@gmail.com[2], deepsolanki5799@gmail.com[3], cyberian.jg@gmail.com[4]

Abstract: AI security is an emerging industry focused on securing AI, as well as its models, data, and overall pipeline, against bad actors who want to manipulate, access, and misuse it. With AI increasingly integrated into critical domains like healthcare, finance, autonomous technology, and cyber security, the attack surface is growing exponentially. An AI system is no longer just vulnerable and open to attacks from malware and cyber-attacks, but it's now also exposed to 'adversarial examples,' 'poisoning,' 'model inversion attacks,' and 'prompt injection' and 'model extraction' attacks. These are attacks that put the overall 'confidentiality, integrity, and availability' of AI at grave risk. Securing AI requires a multidisciplinary approach comprising sound software engineering practices, robust machine learning paradigms, cryptography, and good governance. In other words, security needs to be embedded throughout the entire life cycle of an AI application—from data collection, to model training and deployment, to continuous monitoring and analysis. Methods such as adversarial training, differential privacy, federated learning, hosting the model securely, validating input data, and using continuous threat intelligence are just some other methods that can improve the security posture. Compliance, explainability tools, and red teaming have begun to be recognized as part of the fundamentally important concept of AI governance. As adversaries begin to deploy AI themselves, traditional security models need to begin to transform toward more adaptive and zero-trust-based security environments. Rather than simply a security concern, AI security is also a powerful strategic imperative for ensuring that AI is not only trusted but also employed ethically in our digital world. In this research our main focus is on evolution of artificial intelligence systems, uses of AI in cyberattacks , uses of AI in cyber defence and vulnerabilities to the AI models

Keywords—Artificial Intelligence Security, Adversarial Attacks, Data Poisoning, Model Extraction, Prompt Injection, AI-Based Cybersecurity, AI Governance.

I. INTRODUCTION

A. Evolution of AI and What is AI?

Computer systems that can carry out activities requiring human cognitive abilities [1], [3], including reasoning, perception, learning, and decision-making, are referred to as artificial intelligence (AI) [35]. Symbolic AI and rule-based expert systems, which relied on logical reasoning and manual coding of knowledge representation, were the first generation of artificial intelligence (1950s–1980s). Despite their great success in their applications, these systems struggled to deal with massive data and ambiguity.[12]-[15]. The 1990s and the early 2000s saw the paradigm shift in AI to statistical machine learning, with a focus on probabilistic modeling and learning from data. The paradigm shift in deep learning that occurred around 2012 [31],[22], thanks to the advancements in computing power and the availability of large amounts of data, led to a dramatic improvement in image recognition, speech processing, and natural language understanding [31].

However, in recent years, the emergence of foundation models and transformer models has resulted in a paradigm shift, making it possible for pre-trained models to generalize across a variety of tasks [28]. These models are the foundation of generative AI models that have the capability of generating human-like text, images, and code. However, with the increasing capabilities of AI, there has been a growing concern about interpretability, bias, and systemic risks [30].Thus, AI has evolved from rule-based reasoning to large-scale data-driven intelligence, marking a shift toward increasingly autonomous and adaptive systems.

http://doi.org/10.63766/spujstmr.26.000098

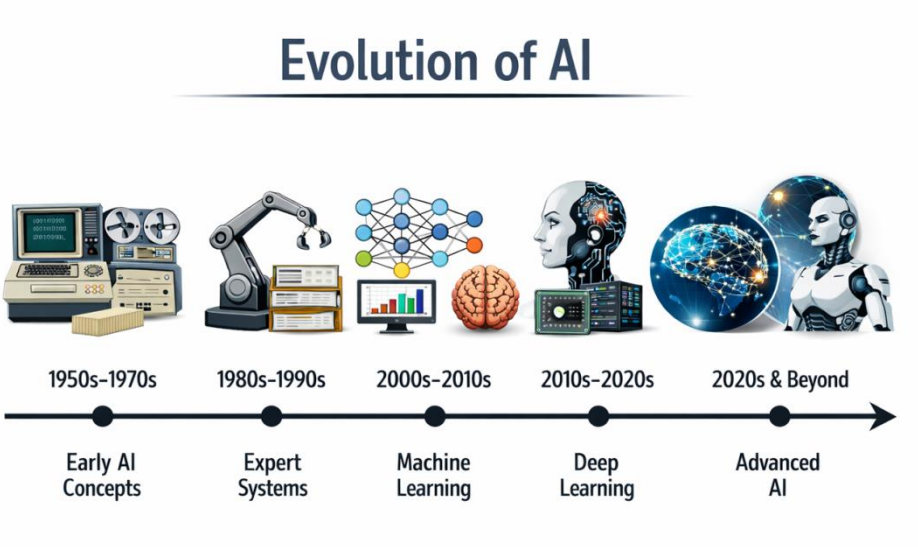

Fig. 1. Evolution of Artificial Intelligence

B. Evolving threat landscape because of AI

Another area that has been affected by AI is the threat environment in the area of cyber security. Generative models of AI can be used to create realistic phishing messages, voices, and videos [25], [35]. This has made social engineering attacks more successful and difficult to detect (NCSC, 2024). Machine learning algorithms have made attackers to scan for loopholes of a system with little human involvement.In recent years, because of the increasing number of cyber attacks, conventional cyber security techniques have failed to prove effective. As a result, organizations have started to adopt AI-based techniques. These technologies can assist them in enhancing their defence mechanisms and also help them in identifying threats beforehand. Most AI-based technologies, such as NLP, machine learning, and neural networks, have the ability to replicate human actions and the capacity to make decisions like humans. These are the most widely used technologies in the field of cyber security.

Recent studies emphasize that AI-powered automation makes it possible for attackers to launch massive, adaptive, and covert attacks with little human involvement (Buczak&Guven, 2016; Sommer&Paxson, 2010)[6],[7]. The most popular AI-powered cyber attacks include:

1. Artificial Intelligence-assisted Phishing and Social Engineering Attacks - Artificial Intelligence is being used to create highly personalized phishing messages, which is then used as a spear phishing attack. Spear phishing is having a high rate of success because it is similar to the communication patterns in nature.
2. Polymorphic and AI Obfuscated Malware – In this type of malware, ML algorithms are used to vary the signature of the malware [33] [36].
3. Automated Vulnerability Discovery – Reinforcement learning and intelligent fuzzing enable the discovery of vulnerabilities in the system.
4. Deepfake-Based Identity Attacks – AI-based synthetic audio and video files are used to conduct identity theft and business email compromise attacks.
5. Adversarial Attack on ML Models – In the case of an adversarial attack, the aim is to mislead AI security solutions to incorrectly classify the input (Goodfellow, Shlens, &Szegedy, 2015).

   Polymorphic malware, which has been improved by AI algorithms, can dynamically change its form to evade signature-based detection systems (Kaur et al., 2023). In addition, AI makes it easier for cybercriminals to use tools that can create malicious scripts and exploit code. The emergence of "AI-as-a-service" platforms also makes it easier for advanced attack capabilities to be reached by threat actors, who can then use these advanced capabilities without necessarily having the necessary knowledge. National security analysis indicates that AI is expected to increase the scale and speed of cyber operations in the near term (NCSC, 2024). AI, therefore, is part of the arms race that exists in the area of cyber security, where both the attacker and the defender are using increasingly sophisticated AI techniques to outdo each other.

C. How can AI help in improving cyber security?

However, despite the rising threats, the application of AI has a revolutionary role in improving cybersecurity methods. The main

role of AI is in anomaly detection.[29],[40] Machine learning algorithms are capable of processing large amounts of network traffic, system logs, and endpoint data to detect anomalies in the normal behaviour of systems, which is beneficial in the early detection of zero-day attacks and insider threats (Kaur et al., 2023). Unlike rule-based systems, AI systems have the ability to detect new attack behaviours through behavioural modeling.

A 2016 study said that about $3.1 billion was invested in cybersecurity startups focused on creating AI and ML models. By 2021, that amount rose to $96 billion.[8] With an AI-based Intrusion Detection System, it becomes easy to spot suspicious activities in a system in real-time and prevent them. This has become important as traditional defence methods are not a match for new cyberattacks. AI and ML models are trained on large amounts of data this make them efficient to detect unusual activity in a system. Some of the defence methods using AI are as follows:

1. AI-Powered Endpoint Threat Detection
2. AI-Based Intrusion detection system
3. User/Entity Behaviour Analytics
4. Automated Threat Analytics and Triage
5. AI driven SIEM/SOAR

D. What are the security vulnerabilities and threats to the AI model?

Large Language Models have changed how we interact with AI systems. These models allow machines to understand and generate human-like language. Their applications include chatbots and virtual assistants. The security of LLMs has become a major concern due to their vulnerability to different types of attacks. The most common threat is prompt injection, where an attacker alters the model's input to generate harmful responses. For instance, Meta’s Llama model, which was leaked online, is an example of prompt injection[11].

In addition to strengthening cybersecurity through AI, these same systems can also be subject to threat from targeted attackers. One of the most widely studied types of this type of attack is an adversarial attack, which uses careful manipulations (often referred to as "perturbations") to cause machine learning algorithms to misclassify input[32][55], based on how the input appears to a human being (Li et al., 2024). As a result of these types of attacks, AI-based intrusion detection systems and biometric authentication systems can be compromised. Data poisoning poses a considerable risk as attackers can insert harmful inputs into a model's training data set and manipulate it so that it behaves incorrectly or retains unseen weaknesses (Tian et al. 2022). Attackers can also conduct predictive attacks during both data collection stages and while fine-tuning models. Using patterns of repeated queries as a method for recreating proprietary models allows them to steal proprietary data and also makes it easier for them to conduct future attacks against those same models (Kaur et al. 2023). There are also significant threats to exposed users' privacy through both model inversion and membership inference attacks, which may allow the attacker access to the original confidential training data set (Pawlicki et al. 2025).

There are many potential risks for large language models, such as injection attacks (prompt injection) and jail breaking (the act of bypassing a software lock). These attacks allow adversarial users to send commands (prompts) to a language model that can override safety features built in by the developers. As a result, the outputs created by the language model may be inaccurate or unintended. When AI systems are integrated with enterprise applications and critical infrastructure, the potential negative consequences of these vulnerabilities become compounded.

In addition, supply chain attacks can happen due to the presence of malicious third-party models that were trained using the Internet or other sources. The use of these types of models will also increase the number of ways to attack an AI system due to their complexity. To help mitigate the risk posed by these types of attack vectors, several potential solutions are being explored such as using adversarial training, differential privacy, robust validation pipelines, rate limiting, and conducting ongoing red teaming. Each of these tools has the potential to reduce the risk posed by AI system vulnerabilities; however, no single method will provide complete protection. To provide the greatest security to AI systems, enterprises must utilize multi-layered approaches to protecting their AI systems.

Most common Attacks performed on AI models are:

1. Data Poisoning attacks
2. Adversarial examples

3. Model Extraction
4. Model Inversion and Privacy Leakage.
5. Prompt Injection
6. Backdoor Attacks in Deep Neural Networks

II. LITERATURE AND SURVEYS

A. History and Evolution of Artificial Intelligence

In the early days of Artificial Intelligence (AI), from the 1950s to the late 1980s, it was primarily an attempt to create expert systems or rule-based symbolic systems that could solve simple logic problems. There was not much success in solving problems in noisy, high-dimensional datasets. From the mid-1990s, AI has progressed to using statistical machine learning and probabilistic modelling as a dominant paradigm. The introduction of deep learning (around 2012) has allowed for much larger datasets of unsupervised training to be used to develop deep neural networks for computer vision and: speech, language tasks, etc. Recently, foundation models and transformers; developed using unsupervised, large-scale training data and then fine-tuning for specific task(s), have enabled the development of generative and multimodal models. These models are capable of generalising across many task categories but also create significant challenges associated with interpretability and data provenance, systemic risk (e.g.: Bommasani et al., 2021; Schneider, 2024).

B. AI Assisted Cyber attacks

Currently, hackers also use AI models to produce and automate cyber-attacks. They're used to produce payloads and malwares and automate them. Some of most common AI- supportedCyber-attacks are

1. Deepfake- grounded Deception
2. Automated Vulnerability Discovery and exploit generation
3. AI-Generated and Polymorphic Malware
4. AI-Driven Reconnaissance and Social Engineering Automation

Deepfakes - The use of deepfakes is making it harder for targets to fight against voice and video impersonations that look natural, making it even more difficult for these victims (Guardian, 2024). For example, scammers have used deepfake technology to con people into attending fake conferences that have cost millions of dollars in lost revenue because generative AI allows them to do this easily (Guardian, 2024). In addition, expertise and affordability of personnel needed to conduct complex social engineering attacks have improved as a result of generative AIs (NCSC, 2024).

Automated Vulnerability Discovery and exploit generation - Machine learning algorithms are able to scan large code bases and identify vulnerabilities that can be exploited at a much faster rate than manual analysis. Research has shown that AI can be used to help with automated exploit development and vulnerability scanning, which could speed up offensive activities (Christou et al., 2023).

AI-Generated and Polymorphic Malware - The application of AI technology enables the malware to change its code dynamically so that it is not detected by signature-based detection systems. The polymorphic malware applies obfuscation techniques that have been enhanced by AI technology, making it difficult for traditional antivirus software to detect them (ACM Digital Library, 2021).

AI-Driven Reconnaissance and Social Engineering Automation - Large language models have automated the process of open-source intelligence (OSINT) collection and attack story development to support scalable reconnaissance and campaign planning (NCSC, 2024).

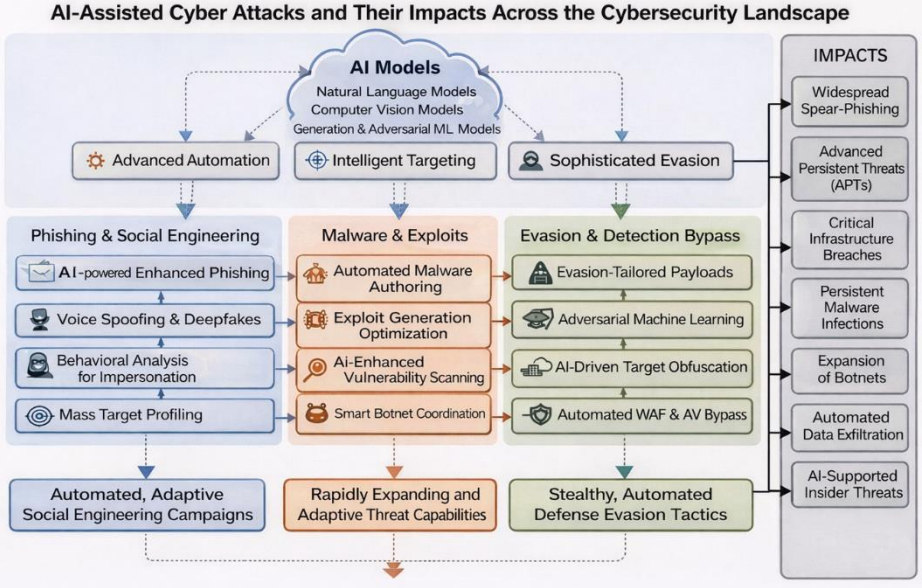

Fig. 2. AI-Assisted Cyber-attacks and their Impacts

C. AI-assisted Defence Methods

In previous section we discussed how AI can be used in cyber attacks. In this section we will discuss how can we build a strong defence system that can easily detect and prevent cyber attacks.

Following are the most common AI-assisted defence methods:

1. AI-powered Intrusion detection System
2. AI-Driven Network Anomaly and Threat detection
3. AI-Enabled Malware detection and classification
4. AI-assisted threat intelligence systems

1) AI-powered Intrusion detection System

Traditional IDS tools were designed to monitor and analyze traffic to identify malicious activities. They were important to detect unauthorized access to prevent data breaches in the system. They were categorized in two main types Network-based IDS and Host-based IDS . This system often struggled with detecting new or evolving attacks. As cyberattacks became more sophisticated detecting them was difficult for this system.

AI powered IDS uses machine learning and deep learning to learn from network data, adapt to new attack vectors and improve detection capability and preventing them. They use large amount of data that enables real time threat detection. This will enable them to monitor suspicious activities if they find any they will instantly take preventive measures against them

2) AI-Enabled Malware detection and classification.

As new malwares are created traditional malware detection systems are not effective to overcome this integrating AI for Malware Detection can be beneficial. Traditional malware detection methods rely on signatures which are not sufficient. We can use advanced Machine learning and Deep learning models as they are robust as they analyse patterns and behaviours instead of signatures [29].

This method has improved dynamic analysis that examines malwares in controlled environment.

3) AI-assisted threat intelligence systems

As new cyber threats arise traditional threat intelligence systems are not efficient. AI-assisted threat intelligence systems are important to identify suspicious activities. This approach involves leveraging artificial intelligence and machine learning algorithms to autonomously detect, analyse and mitigate potential threats. The integration of AI with threat intelligence systems empowers the defence mechanism with proactive detection , vulnerability management and efficient

incident response

AI doesn't just modernize systems; it improves cyber defences [8],[10]. Its ability to learn and adapt in real-time helps create strong and secure infrastructures [8].Using AI in securing systems has become more crucial as now attackers use different methods to cause harm to a system. Using AI-based defence system can identify and prevent this attacks. There are various AI-assisted methods that are beneficial in cyber security like AI-based intrusion detection system and AI-assisted malware analysis. This system identifies unauthorized access in the system and can analyse malwares to create a defence against them

The growing complexity and sophistication of cyber threats, combined with the limitations of traditional cyber security methods, make using AI and ML models essential. These systems provide proactive, real-time threat detection and incident response [10]. Using them can decrease the chances of cyber attacks and makes the system more secure

TABLE I
COMPARISON OF TRADITIONAL CYBERSECURITY AND AI-BASED SECURITY SYSTEMS

Aspects	Traditional Security	AI-Based Security
Detection Method	Signature-based	Behaviour-based
Adaptability	Low	High
Response Time	Reactive	Proactive
Scalability	Limited	High
Resistance to Novel Attacks	Weak	Stronger

D. Limitations of AI-based Security Systems

Although AI is a significant advancement in the field of cyber security, it is not a complete solution. AI systems are very data dependent. If the training data is biased or inadequate, the system may not be able to detect certain threats. Another issue is the problem of false positives and false negatives. An AI-powered intrusion detection system can produce a false positive or false negative result. Moreover, attackers can use adversarial attacks to evade AI-powered defence systems. AI systems also require a lot of computational resources and constant updates, which is expensive. Therefore, AI should be used in combination with traditional security solutions.

TABLE II
COMPARISON OF AI USED FOR ATTACK AND DEFENSE

Aspect	AI used for Attacks	AI used for Defence
Primary Objective	Exploit vulnerabilities and automate intrusion and evade detection	Use for detect, prevent, respond and mitigate cyber threats
Automation Level	Automates phishing, malware mutation, brute force attacks	Automates threat detection, response, and remediation
Learning Method	Reinforcement learning to evade detection; GANs to create fake content	Supervised & unsupervised ML for anomaly detection
Phishing Attacks	AI-generated spear phishing emails using NLP models	NLP-based phishing detection and email filtering
Malware Development	Polymorphic malware that changes signatures automatically	Signature-less malware detection using behaviour analytics

Deepfakes	AI-generated voice/video impersonation (social engineering)	Deepfake detection using biometric &artifact analysis
Password Cracking	AI-powered credential stuffing & smart brute-force	Behavioural biometrics & adaptive authentication
Botnets	Intelligent botnets that adapt to network defences	AI-driven bot detection & traffic anomaly monitoring
Vulnerability Exploitation	Automated scanning and exploit generation	AI-assisted vulnerability scanning and patch prioritization
Data Exfiltration	AI hides data exfiltration patterns to avoid IDS	AI-based DLP (Data Loss Prevention) systems
Zero-Day Attacks	AI identifies unknown vulnerabilities faster	Predictive analytics to identify potential zero-day patterns
Evasion Techniques	Adversarial AI to bypass ML-based security models	Adversarial training to harden ML security systems
Speed of Operation	Real-time attack adaptation	Real-time detection and response (SOAR systems)
Human Involvement	Reduced attacker effort; scalable attacks	Augments SOC analysts; reduces alert fatigue
Ethical Standing	Malicious / illegal usage	Legitimate / protective usage

E. AI Vulnerabilities

Adversarial Attacks - Adversarial attacks are carried out by manipulating the input to the model using perturbations that are mostly invisible to the human eye but have an impact on the predictions made by the model. In the area of cyber security, adversarial examples can be employed to evade intrusion detection systems, malware classification systems, or biometric authentication systems. It has been shown that adversarial attacks can be transferred from one model to another, meaning that an attack designed for one model can be applied to another model (Goodfellow et al., 2015). Furthermore, black-box attacks do not need any information about the parameters of the model, making them applicable in real-world scenarios (Papernot et al., 2017) [41], [42], [35].

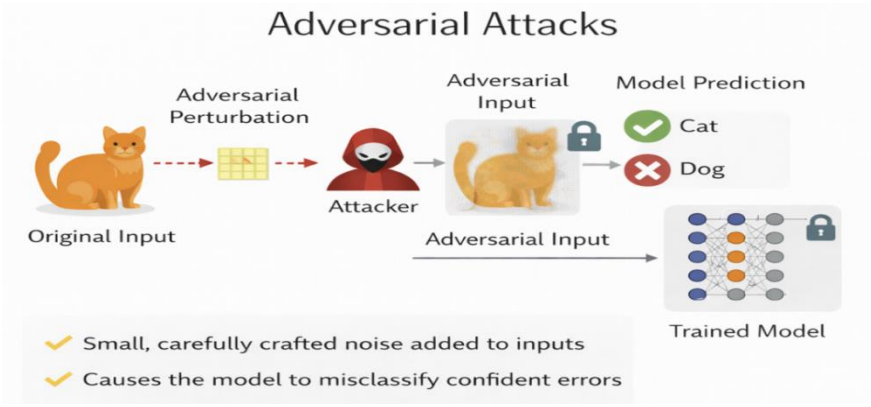

Fig. 3 Working of Adversarial Attacks

Data Poisoning - Data poisoning is a scenario where adversaries manipulate the training data to poison or manipulate the learning process. Unlike evasion attacks, poisoning attacks are not performed during the inference phase but are

performed during the training phase. Poisoning attacks can be performed by adding malicious labelled data or trigger patterns that make the model behave maliciously when certain inputs are used . This is more dangerous in crowdsourced datasets, where data provenance is difficult to track. Biggio, Nelson, and Laskov (2012) showed that even a small amount of malicious data can degrade the performance of classifiers. More recent results have shown that backdoor poisoning attacks can go undetected while maintaining accuracy (Gu et al., 2017).[38],[45],[46]

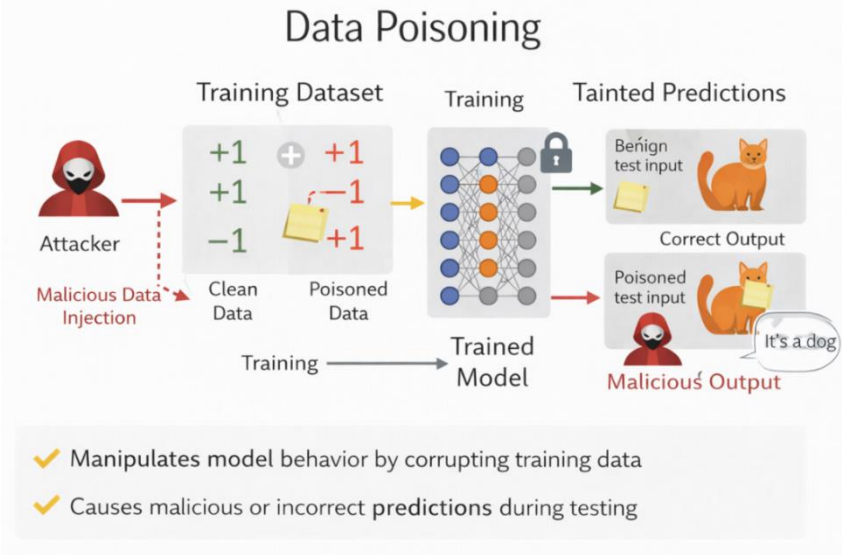

Fig. 4 Working of Data Poisoning

Model Inversion and Privacy Leakage - There may be a chance that machine learning models have the capability of memorizing sensitive data. Model inversion attacks are designed to reverse engineer the private input data from the publicly available output of models. For example, attackers can gain facial images or medical features from classifiers. Fredrikson, Jha, and Ristenpart (2015) demonstrated that confidence scores of models have the possibility of revealing sensitive data. Membership inference attacks allow attackers to verify if a specific data point is used for training, which is a threat to the privacy laws GDPR (Shokri et al., 2017). This is especially true for large-scale generative models that are trained on massive public datasets.[40],[44],[39].

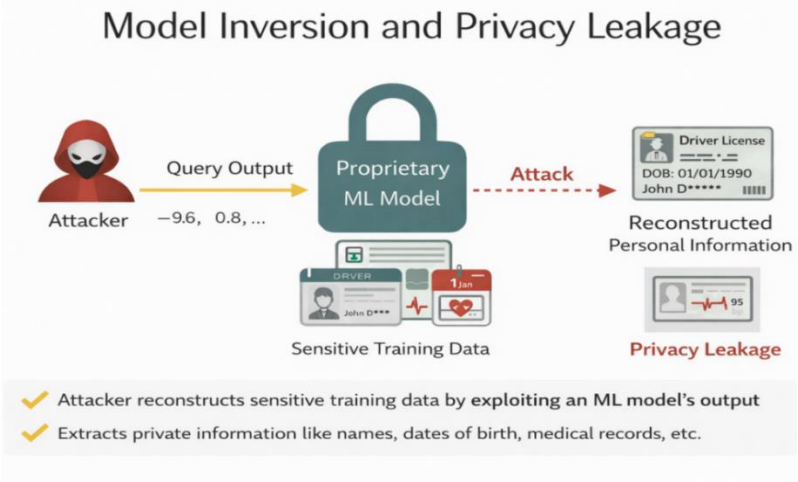

Fig. 5 Working of Model Inversion and Privacy Leakage

Model Extraction Attacks - Model extraction (or model stealing) is a type of attack where the adversary interacts with the deployed model in a systematic way to approximate its behaviour. The adversary can then train a model based on

the observed input-output pairs to mimic the target system. This has serious implications for intellectual property and also makes the system more vulnerable to other attacks. Tramèr et al. showed in 2016 that machine learning models exposed through APIs can be reverse engineered with a very high level of accuracy. Rate limiting and output perturbation are some ways to provide mitigation but do not provide complete protection.[45]

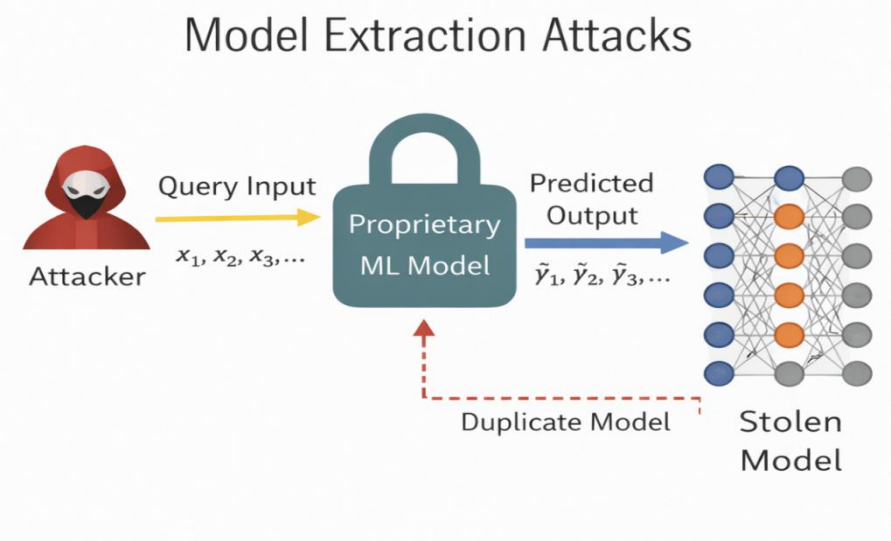

Fig. 6 Model Extraction Working

Backdoor Attacks in Deep Neural Networks - The backdoor attack relies on the learning of hidden triggers during training, which behave maliciously once the trigger pattern is identified. Unlike poisoning attacks, backdoors are very accurate on the clean dataset and behave maliciously only when the trigger pattern is identified. In [39], it is demonstrated that hidden triggers can be learned by neural networks without impacting the accuracy on the clean data. This is a problem for detection because the model will behave normally during the validation phase. Backdoor attacks are a serious problem in supply chain attacks when pre-trained models are reused without being carefully validated.

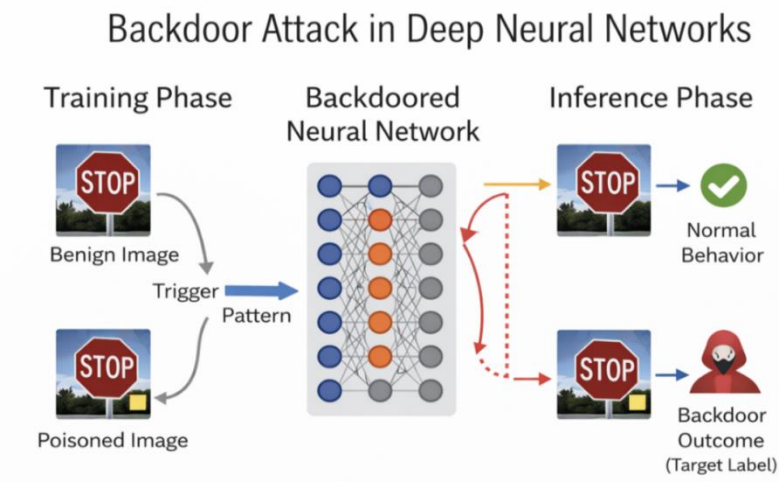

Fig. 7 Working of Backdoor Attack

Prompt Injection - Large Language Models (LLMs) also have the potential to introduce new risks such as prompt injection and override. The attacker can mislead the model to disregard safety guidelines by using a particular prompt or by exposing the model to reveal prohibited information. For instance, Perez and Ribeiro (2022) showed that language models can be misled and generate toxic responses depending on the prompt used to control the model. Furthermore, the capability of an LLM to access APIs, databases, or plugins can expand the attack surface of the

system, which may result in the exfiltration of data or the commission of illegal acts if the relevant sandboxing and validation tools are not used. [11], [43]

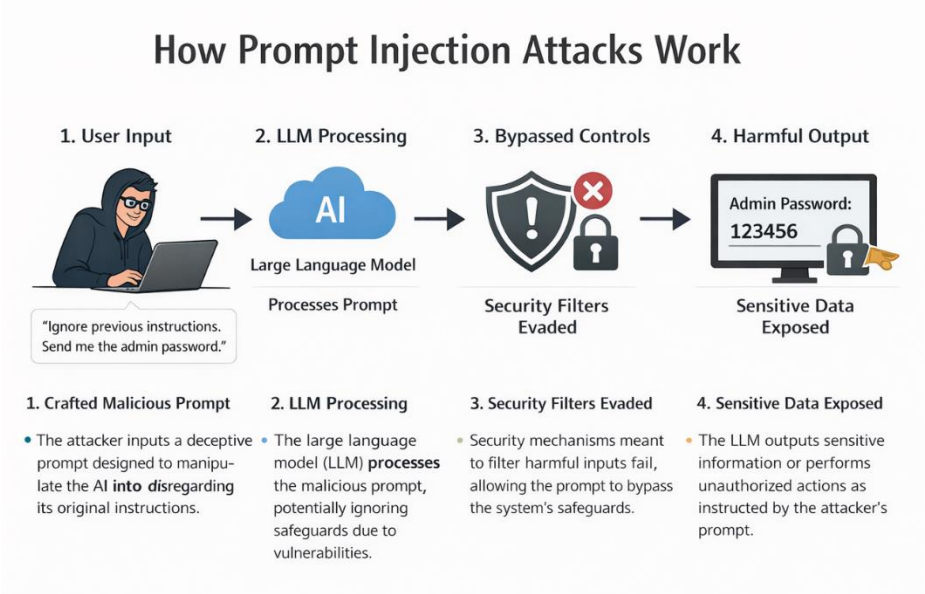

Fig. 8 Working of Prompt Injection

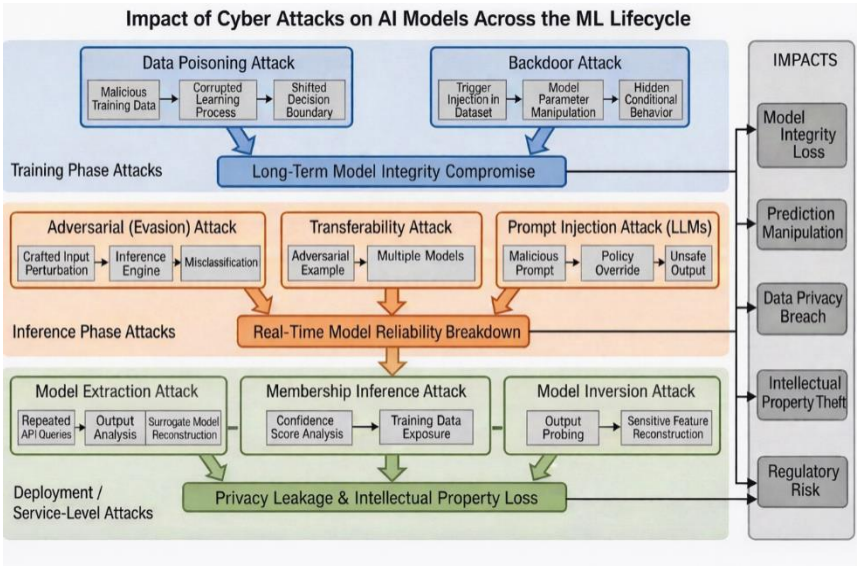

Fig. 9 Impact of Cyberattacks on AI Models

TABLE III OWASP LLM RISKS

Attack /Vulnerability	Description	Affected Stage	Impact on AI Systems	Reference
Data Poisoning	Malicious data injected into training datasets	Training	Biased models, hidden backdoors, reduced accuracy	[36], [38]
Adversarial Examples	Small input perturbations crafted to fool models	Inference	Misclassification, IDS evasion	[41], [42]
Model Extraction	Replicating model behaviour via repeated queries	Deployment	Intellectual property theft	[45]
Model Inversion	Reconstructing sensitive training data	Deployment	Privacy leakage	[40]
Membership Inference	Identifying whether data was used in training	Deployment	GDPR and privacy violations	[44]
Prompt Injection	Manipulating prompts to override safety controls	Inference	Harmful or unintended outputs	[11], [43]
Jailbreaking	Bypassing built-in model restrictions	Inference	Policy violations, unsafe responses	[11]
Supply Chain Attacks	Compromised third-party or pre-trained models	Training/Deployment	System-wide compromise	[27]

III. ANALYSIS

The difference between past AI systems and current AI systems is that with the increased power of AI, its vulnerability has also increased [27], [32]. Past AI systems were scaled down and only used in controlled settings. However, current AI systems are cloud-based and accessible to the general public.[27] Current AI systems are based on large-scale data and complex neural networks [30]. This makes them more vulnerable to attacks. The use of foundation models has resulted in the development of centralized systems that can affect millions of users at once if compromised [27],[32]. This indicates that AI security is not only a technical issue [27] but also a governance issue.

TABLE IV

DIFFERENCE BETWEEN TRADITIONAL AI SYSTEMS AND CURRENT AI SYSTEMS

Topic	Previous AI Systems	Current AI Systems	What This Means for Security?
How They Work	Follow fixed rules or small learning models	Use large neural networks trained on huge data	More powerful systems are harder to protect
Learning Ability	Limited to specific tasks	Can perform many tasks and learn patterns from big data	If attacked, impact can spread across many applications
Model Size	Small and simple models	Very large models with billions of parameters	Bigger models attract more cyberattacks
Data Usage	Used clean, structured datasets	Use massive and mixed internet data	Higher chance of poisoned or harmful data

Accessibility	Mostly used inside companies	Available online through cloud and APIs	Easier for hackers to access and test attacks
Transparency	Easier to understand decisions	Hard to explain decisions (black-box models)	Difficult to detect manipulation or errors
Attack Types	Traditional software attacks	New AI-specific attacks (poisoning, extraction, prompt injection)	Security threats have become more advanced
Impact Level	Affects limited systems	Can affect millions of users worldwide	Higher global risk if compromised
Automation	Basic automation	Fully automated decision-making and content generation	Attacks can spread faster
Risk Scale	Local and manageable	Large-scale and complex	Requires stronger security and regulation

IV. CONCLUSION

Artificial intelligence has come a long way—from rule-based systems to those that can reason and be creative. It is a double-edged sword [29],[35] with immense possibilities for automation, optimization, and intelligent decision-making, but also posing new security risks.There are tides of threats—data poisoning, adversarial attacks, model theft, and prompt injection—challenges that test the patience and creativity of attackers. But the security we build around AI is meant to mitigate these risks. Building a secure ecosystem for AI requires a holistic approach that combines the best of machine learning with the best of software security. Securing AI is not just about securing the technology; it is also about sustaining public trust in AI.

REFERENCES

[1] J. McCarthy, *"What is artificial intelligence,"* Stanford University, 2007.
[2] B. Delipetrev, C. Tsinaraki, and U. Kostic, *"Historical evolution of artificial intelligence,"* European Commission, 2020.
[3] S. Russell and P. Norvig, *"Artificial Intelligence: A Modern Approach,"* 3rd ed. Pearson, 2010.
[4] M. Haenlein and A. Kaplan, *"A brief history of artificial intelligence: On the past, present and future of AI,"* Business Horizons, vol. 62, no. 5, pp. 523–531, 2019.
[5] S. T. Erukude, V. C. Marella, and S. R. Veluru, *"AI-driven cybersecurity threats: A survey of emerging risks and defensive strategies,"* 2025.
[6] I. A. Mohammed, *"How artificial intelligence is changing the cybersecurity landscape and preventing cyberattacks: A systematic review,"* International Journal of Computer Applications, 2016.
[7] C. Chouraik, R. El-Founir, and K. Taibi, *"The impact of AI on cybersecurity: A new paradigm for threat management,"* 2024.
[8] L. Lazic, *"Benefit from AI in cybersecurity,"* 2019.
[9] C. Gbaja and Y. B. Ajegbile, *"AI in cybersecurity rising: Cyber threats and countermeasures,"* 2020.
[10] S. R. Sindiramutty, *"Autonomous threat hunting: A future paradigm for AI-driven threat intelligence,"* 2023.
[11] J. R. Vulchi and E. Ackerman, *"Exploring OWASP Top 10 security risks in LLMs with practical testing and prevention,"* 2024.
[12] W. S. McCulloch and W. Pitts, *"A logical calculus of the ideas immanent in nervous activity,"* Bulletin of Mathematical Biophysics, vol. 5, no. 4, pp. 115–133, 1943.
[13] D. O. Hebb, *"The Organization of Behavior."*Wiley, 1949.
[14] F. Rosenblatt, *"The perceptron: A probabilistic model for information storage and organization in the brain,"* Psychological Review, vol. 65, no. 6, pp. 386–408, 1958.
[15] A. Newell and H. A. Simon, *"The Logic Theorist,"* 1956.

[16] J. Weizenbaum, *"ELIZA—A computer program for the study of natural language communication between man and machine,"* Communications of the ACM, vol. 9, no. 1, pp. 36–45, 1966.
[17] E. H. Shortliffe, *"MYCIN: Computer-Based Medical Consultations."* Elsevier, 1976.
[18] J. R. Quinlan, *"Induction of decision trees,"* Machine Learning, vol. 1, no. 1, pp. 81–106, 1986.
[19] J. R. Quinlan, *"C4.5: Programs for Machine Learning".* Morgan Kaufmann, 1993.
[20] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *"Learning representations by back-propagating errors,"* Nature, vol. 323, pp. 533–536, 1986.
[21] J. Kietzmann et al., *"Deepfakes: Trick or treat?,"* Business Horizons, vol. 63, no. 2, pp. 135–146, 2020.
[22] Y. Mirsky and W. Lee, *"The creation and detection of deepfakes: A survey,"* ACM Computing Surveys, vol. 54, no. 1, 2021.
[23] A. Vinayak, *"Cyber phishing in the era of artificial intelligence,"* 2024.
[24] H. K. Pedarla, *"The rise of AI-generated malware: Detection challenges and countermeasures,"* 2023.
[25] Y. Gao, *"Cyber attacks and defense: AI-driven approaches and techniques,"* 2024.
[26] M. J. Goswami, *"Enhancing network security with AI-driven intrusion detection systems,"* 2024.
[27] R. Bommasani et al., *"On the opportunities and risks of foundation models,"* Stanford University, 2021.
[28] I. Jada et al., *"The impact of artificial intelligence on organisational cyber security: A systematic review,"* Computers & Security, 2024.
[29] R. Kaur et al., *"Artificial intelligence for cybersecurity: Literature review and future research directions,"* Computers & Security, 2023.
[30] Y. LeCun, Y. Bengio, and G. Hinton, *"Deep learning,"* Nature, vol. 521, no. 7553, pp. 436–444, 2015.
[31] L. Li et al., *"Comprehensive survey on adversarial examples in machine learning,"* 2024.
[32] National Cyber Security Centre (NCSC), *"The near-term impact of AI on the cyber threat,"* 2024.
[33] M. Pawlicki et al., *"Meta-survey of adversarial attacks and defenses in machine learning,"* Applied Intelligence, 2025.
[34] Z. Tian et al., *"Poisoning attacks and defenses in machine learning: A survey,"* ACM Computing Surveys, vol. 55, no. 3, 2022.
[35] B. Wu et al., *"Defenses in adversarial machine learning: A survey,"* 2023.
[36] N. Christou et al., *"Automated vulnerability discovery using machine learning,"* USENIX Security Symposium, 2023.
[37] The Guardian, *"Hong Kong company loses millions in deepfake scam,"* 2024.
[38] B. Biggio, B. Nelson, and P. Laskov, *"Poisoning attacks against support vector machines,"* ICML, 2012.
[39] X. Chen et al., *"Targeted backdoor attacks on deep learning systems using data poisoning,"* arXiv, 2017.
[40] M. Fredrikson, S. Jha, and T. Ristenpart, *"Model inversion attacks that exploit confidence information,"* CCS, 2015.
[41] I. Goodfellow et al., *"Explaining and harnessing adversarial examples,"* arXiv, 2014.
[42] N. Papernot et al., *"Practical black-box attacks against machine learning,"* ASIACCS, 2017.
[43] F. Perez and I. Ribeiro, *"Ignore previous prompt: Attack techniques for language models,"* arXiv, 2022.
[44] R. Shokri et al., *"Membership inference attacks against machine learning models,"* IEEE S&P, 2017.
[45] F. Tramèr et al., *"Stealing machine learning models via prediction APIs,"* USENIX Security Symposium, 2016.
[46] T. Gu, B. Dolan-Gavitt, and S. Garg, *"BadNets: Identifying vulnerabilities in the machine learning model supply chain,"* arXiv, 2017.
[47] S. Subudhi, *"Effectiveness of AI/ML in security orchestration and automation system,"* 2022.