Scientific Research Publishing

Journal of Information Security, 2026, 17(1), 1-18
https://www.scirp.org/journal/jis
ISSN Online: 2153-1242
ISSN Print: 2153-1234

Artificial Intelligence Systems Cybersecurity Ensuring: Analysis of Vulnerabilities, Attacks, and Countermeasures

Conrad Onesime Oboulhas Tsahat, Ngoulou-A-Ndzieli

Ecole Nationale Supérieure Polytechnique, Université Marien Ngouabi, Brazzaville, Republic of Congo
Email: oboulhas@yahoo.fr, becker20000@yahoo.fr

How to cite this paper: Oboulhas Tsahat, C.O. and Ngoulou-A-Ndzieli (2026) Artificial Intelligence Systems Cybersecurity Ensuring: Analysis of Vulnerabilities, Attacks, and Countermeasures. *Journal of Information Security*, **17**, 1-18.
https://doi.org/10.4236/jis.2026.171001

Received: September 8, 2025
Accepted: December 19, 2025
Published: December 22, 2025

Abstract

The rapid adoption of Artificial Intelligence (AI) systems in critical sectors of society has given rise to new cybersecurity challenges. Unlike traditional software systems, AI systems have unique characteristics such as data dependence, model complexity, and adaptive behavior, which create new types of vulnerabilities and attack vectors. Through such attacks, intruders can manipulate these systems to change their behavior to achieve their goals. According to expert data, only 25% of modern artificial intelligence applications are properly protected. Given these technologies specifics their security covers a wide range of tasks, including the data protection, algorithmic models and application scenarios. This review article provides a comprehensive analysis of the current state of AI cybersecurity, systematizing the vulnerabilities inherent in AI, classifying the main types of attacks at different stages of the AI lifecycle, and describing adequate countermeasures. This paper proposes a comprehensive taxonomy of threats and defenses, covering aspects from data collection to model deployment and operation. The goal of the paper is to provide a deep understanding of the complex AI threat landscape and guide researchers and practitioners in the development and implementation of robust and secure AI systems. Finally, current research gaps are identified and future directions are outlined to ensure the sustainability of AI in a dynamically changing digital environment.

Keywords

Artificial Intelligence, Cybersecurity, Vulnerability, Attacks, Machine Learning

1. Introduction

Artificial intelligence (AI) systems have permeated every modern society aspect

performing a functions variety from driving vehicles and helping doctors diagnose diseases to interacting with customers as online chatbots [1] [2]. This remarkable progress is primarily due to significant breakthroughs in machine learning (ML), deep learning, and the enormous increase in computing power [3]-[5]. However, even as AI technology has advanced and become more sophisticated it remains highly vulnerable to an ever-increasing number of cybersecurity threats and malicious attacks. Hackers can deliberately confuse or even "infect" AI systems to cause them to crash, and developers currently have no reliable means of protecting against this. Unlike traditional software, AI systems possess unique characteristics that significantly expand their attack surface and introduce fundamentally new threat vectors. These characteristics include a strong dependence on the quality and integrity of training data, the complexity and opacity of many models ("black box"), and their adaptive nature [6]-[8]. Attackers can exploit these features to manipulate the behavior of AI systems, extract sensitive information, or disrupt their normal functioning, potentially causing significant financial losses, reputational damage, and, in some cases, a threat to human safety [9]. Expert assessments and reports highlight the urgency of the issue, showing that the proportion of properly secured AI applications remains low, and many organizations face AI security challenges, including vulnerability management and maintaining regulatory compliance [10].

The following statistics clearly demonstrate the alarming and significantnature of this trend. The number of incidents related to cyber attacks on AI systems in 2025 was about 16,200 cases, which is 49% more than in 2024. The average cost of an incident reached $ 4.8 million in 2025 [11] [12]. **Table 1** provides major incidents brief description involving attacks on AI over the past 3 years and their consequences.

Table 1. Description of attacks on AI systems and their consequences.

Year	Incident type and description	Consequences nature	Damage assessment	Notes
2023	Samsung leaks information via public AI chatbots	Disclosure of confidential source code and technical documents	Estimated at several billion dollars	Internal data transferred to the system without encryption. Access to AI services within the company is restricted
	Deepfake Voice Synthesis Scam	Fake CEO's voice, deceive the banking system	Direct loss of $18.5 million	The financial institution transferred funds to false details
2024	Fines imposed for privacy violations when working with AI chatbots	Leakage of chat history, accounts, financial information	15 million euros	Sanctions from the European Data Protection Supervisory Authority
	Fake images and videos mass distribution (deepfake)	Manipulation, reputational losses, financial fraud	Total global damage is more than $1.2 billion	More than 105,000 incidents were recorded during the year
	AI-powered ransomware attacks	Blocking of production and corporate systems, data loss	Global losses are more than $20 million	The average ransom was $1.8 million per incident
2025	Targeted Attack on Corporate AI Assistants (Microsoft Copilot)	Unauthorized access to documents, emails and cloud data	Potentially affecting millions of users	Vulnerability in the module for generating responses based on corporate databases

Continued

Massive security breaches in systems with AI components	Leaks of user and service data, reputational losses	Average damage estimate: $4.8 million per incident	Nearly 16,000 serious cases reported, up 49% from 2024
Financial fines and penalties for misuse of AI technologies	Personal data disclosure during models training, subjects rights violation	The average fine in the sector is $35.2 million	US and EU regulators tighten oversight of AI data processing
Leak from a major AI service (DeepSeek, South Korea)	Logs disclosure, dialog contents, API keys and internal environment variables	The financial implications have not yet been disclosed.	Discovered by Wiz Research

AI system components include data, a machine learning model, the infrastructure and processes needed to use it. Because machine learning approaches rely on large amounts of data, AI systems face additional security and privacy concerns beyond classic cyber threats. Additionally, AI tools are increasingly being connected to corporate documents and databases for specific use cases. This integration into existing systems only increases the attack surface exposing organizations to the threat of attackers gaining access to confidential and proprietary information.

The purpose of this review article is to conduct a comprehensive and systematic analysis of the AI cybersecurity landscape, identify key vulnerabilities, classify the main attack vectors, and propose effective countermeasure strategies throughout the AI lifecycle.

To achieve this goal, the following objectives were set to:

- systematize and classify the main types of vulnerabilities inherent in AI systems at various stages of their design, development, and deployment;
- develop a comprehensive taxonomy of cyberattacks on AI systems, detailing the mechanisms, targets, and potential consequences of attacks on data, algorithms, and platforms;
- identify and describe the most effective technical, organizational, and regulatory countermeasures and risk mitigation strategies;
- identify existing gaps in current research and practical approaches to ensuring AI cybersecurity, and outline promising areas for future research.

The scientific novelty of this work lies in the presentation of a comprehensive, multi-layered framework for the analysis and mitigation of cybersecurity risks of AI systems, integrating vulnerabilities, attacks, and countermeasures at each stage of the AI life cycle.

Unlike existing reviews, which often focus on individual aspects, such as adversarial attacks or data protection, this paper offers a holistic view that covers a unique classification of vulnerabilities linked to the phases of the AI lifecycle (see Table 2), allowing for more targeted application of protective measures. It also presents an expanded attack taxonomy (see Table 3) that categorizes threats not only by target (data, algorithm, platform) but also by their mechanism and deployment location, providing a more detailed understanding of impact vectors.

Furthermore, the paper provides an integrated set of countermeasures pro-

posed in the context of the identified vulnerabilities and attacks, facilitating the development of comprehensive defense strategies rather than isolated solutions. This framework serves as a foundation for a more structured and proactive approach to ensuring the security of AI systems, providing researchers and practitioners with a systematic knowledge base for analysis, risk assessment, and the development of effective defense mechanisms.

2. Literature Review

Advancements in Data Science and Computer Science have led to the emergence of ML, the most prominent type of AI in organisational cyber security [13] [14]. The vast amounts of data generated by organisations provide opportunities for a wide range of ML applications in cyberspace, including threat intelligence, anomaly detection, and automation of cybersecurity-related tasks [15]. ML's role in cybersecurity dates to the 1990s with the development of anomaly detection systems (ADS) and intrusion detection systems (IDS) [16], though progress was hindered by data and computing limitations [17]. Today, AI is integral to cybersecurity, transcending corporate jargon [18] [19]. It can simulate human intelligence and behaviours, resulting in automation in cyber security beyond human capability, which can detect a security breach in a network within seconds [20]. General information about AI systems, their cybersecurity state, as well as the classification and attack mechanisms description on them are presented in many different works [21]-[24]. Abuse attacks that are typical for generative AI used by intruders in the process of manipulating tools such as chatbots and image generation tools are described in detail by Chen Yang [25], Garbuk S.V. [26], Jie Yang, Jun Zheng [27].

3. Methodology

The article purpose to consider AI systems cybersecurity ensuring features taking into account their vulnerabilities, possible attacks and countermeasures. Research methodology combines system and comparative analysis, modeling, forecasting, grouping, scenario assessment.

The term "AI cybersecurity" refers to a broad range of factors, strategies, technologies and regulatory measures aimed at protecting AI systems from cyber risks, threats and malicious attacks. It includes identifying, assessing and addressing potential vulnerabilities in AI systems, as well as developing specific measures to protect these systems from malicious actors [23] [28] [29]. Assessing cyber risks in the evolving AI landscape is challenging due to AI growing complexity and interconnectedness with many vital systems. Research institutions, government agencies and leading AI organizations are actively exploring effective approaches to address these challenges.

3.1. Classification of AI System Vulnerabilities

Vulnerabilities in AI systems differ significantly from those found in traditional

software, although some classic vulnerabilities, such as those in infrastructure code, remain relevant. The unique nature of AI, driven by data and models, expands the attack surface and requires a systematic approach to risk identification. These vulnerabilities can manifest themselves at various stages of the AI lifecycle, from design to deployment and operation. Table 2 provides an overview of the vulnerabilities and attack types specific to each stage of the AI lifecycle.

Table 2. AI life cycle, vulnerabilities, and attacks.

AI Life Cycle Stage	Characteristic of vulnerabilities	Attack types of attack (most common)
Design	Lack of a robust security architecture (inadequate threat modeling, weak privacy guarantees, insecure authentication); Bias in data/algorithm design; Weak resilience to adversarial attacks; Lack of explainability (Explainability)	No direct attacks at this stage, but the foundations for future attacks are laid (e.g. Data Poisoning, Adversarial Examples)
Data Acquisition and Preparation	Insufficient data validation and sanitization; Unprotected data storage and transmission; Bias in the source data	Data Poisoning; Data Leakage; Inference Attacks
Model Development and Training	Insecure AI code (vulnerabilities in libraries, frameworks, and AI-generated code); Insecure AI supply chain (compromise of pre-trained models); Weak model robustness to new data	Data Poisoning; Model Stealing/Extraction; Model Inversion; Backdoor Attacks
Deployment	Vulnerable API endpoints; Cloud service configuration errors; Lack of data encryption in transit; Insecure runtimes	API Exploitation; Runtime Exploits; Denial of Service (DoS/DDoS)
Monitoring and Maintenance	Model vulnerability to adversarial perturbations; Insufficient monitoring of model performance and behavior; Vulnerability to "unauthorized" behavior of AI agents	Adversarial Examples; Prompt Injection; Denial of Service (DoS/DDoS); Chained Prompt Injection; Inference Attacks

- *Design Phase Vulnerabilities*

The design phase of an AI system lays the architectural and methodological foundations that can become a source of fundamental vulnerabilities if security aspects are not taken into account. These include the lack of a robust security architecture, which manifests itself in inadequate threat modeling, insufficient data privacy guarantees, and insecure authentication and authorization. Insufficient early assessment of potential attacks and attackers leads to a lack of built-in security mechanisms, while the lack of clear mechanisms for protecting confidential information and controlling access to AI components creates vulnerabilities. Also, if the possible presence of bias in data or algorithms is not taken into account at the design stage, this may lead to unfair, discriminatory, or inaccurate model results, representing both an ethical and a security vulnerability. Additionally, vulnerabilities in system and model design include weak resilience to adversarial attacks due to the design of models without taking into account robustness enhancement mechanisms.

The lack of explainability mechanisms makes it difficult to detect and diagnose attacks or incorrect behavior, and insufficient scalability and resilience to DoS attacks arise if the infrastructure does not take into account protection against massive overloads.

- *Development Phase Vulnerabilities*

The development phase involves coding, model training, and component integration, during which vulnerabilities can be either classic software vulnerabilities

or AI-specific ones. Insecure AI code and code vulnerabilities include risks associated with the use of AI tools to generate code, which may unintentionally introduce vulnerabilities such as SQL injection or XSS if the model is trained on incorrect data or does not undergo rigorous security audits. Also significant is the reliance on third-party libraries and frameworks, which can introduce known or unknown vulnerabilities. Failure to adhere to secure development standards when writing code for AI systems, such as lack of input validation or insecure error handling, is also a critical factor. Insecure data processing in AI systems includes inadequate data validation and sanitization, leading to vulnerability to data poisoning. AI involves inadequate data validation and sanitization, leading to vulnerability to data poisoning. Data poisoning is the intentional introduction of malicious or incorrect data into a training set, leading to a model being trained on erroneous or biased patterns.

Furthermore, inadequate data security during model training can lead to data compromise. An insecure AI supply chain is also a source of vulnerabilities, with attackers compromising data or models during the delivery phase by injecting malicious code or data into components supplied by external suppliers.

- *Deployment and maintenance Phase Vulnerabilities*

Once an AI system is deployed in a production environment, vulnerabilities arise related to configuration, interaction with the external environment, and continuous monitoring. These include vulnerable API endpoints due to insecure configuration or lack of proper authentication and authorization for the application programming interfaces (APIs) through which interaction with the AI model occurs. There's also the issue of a lack of encryption when transmitting sensitive data between AI components or with users. Configuration errors in cloud services hosting AI systems can lead to unauthorized access to data or models. Model theft (Extraction) and data exfiltration (Inference) are serious threats that allow attackers to gain access to an AI model itself, extract its internal architecture and parameters, or reconstruct sensitive information from it. Prompt injection vulnerabilities pose a particular threat to generative models (LLM), as they allow model behavior to be manipulated by injecting malicious instructions through user input and bypassing internal protection mechanisms. Insufficient monitoring of model performance, detection of data drift or anomalous behavior, and delays in applying security patches are also critical vulnerabilities at this stage. Understanding these vulnerabilities at different stages of the AI lifecycle is critical to developing effective protection and risk mitigation strategies.

3.2. Taxonomy of Cyber Attacks on AI Systems

Cyberattacks on AI systems represent a complex and multifaceted threat landscape that differs significantly from attacks on traditional information systems. These attacks exploit unique vulnerabilities inherent in AI models, data, and infrastructure to manipulate system behavior, compromise its integrity, or compromise confidentiality. Currently available expert reports and scientific publications

analysis allowed the author to classify possible attacks on AI systems. The proposed approach is presented in Figure 1.

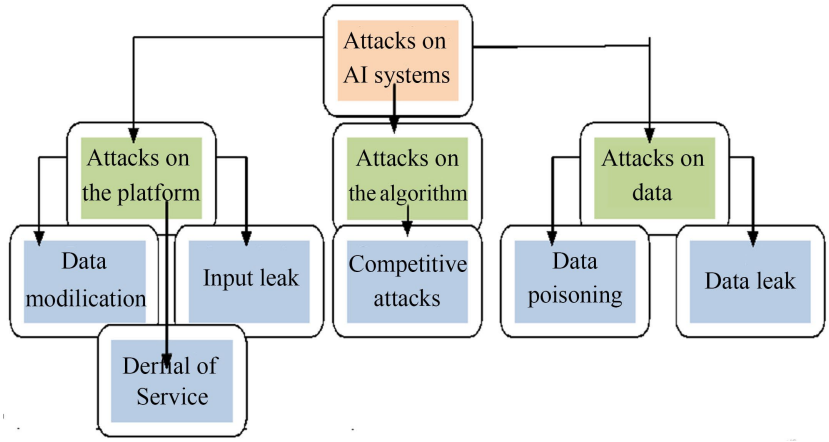

Figure 1. Classification of attacks on AI systems.

To better understand these threats, a systematic taxonomy is proposed that categorizes attacks by their primary target: data attacks, algorithm/model attacks, and platform/infrastructure attacks (Table 3).

Table 3. Taxonomy of cyber attacks on AI systems and examples of countermeasures.

Attack Type	Target (object of influence)	Method	Main Consequences	Examples of Countermeasures
Data poisoning	Data, Model	Injecting malicious, incorrectly labeled data into the training set	Distortion of model behavior, reduction in accuracy, introduction of backdoors, forecast bias	Robust input data validation and sanitization; Outlier detection; Data provenance audit; Use of trusted data sources; Robust learning algorithms; Federated learning.
Adversarial examples	Data, Model	Adding human-imperceptible perturbations to input data	Misclassification/model prediction; Bypassing detection systems (e.g., spam filters, antiviruses)	Adversarial Training; Robust Optimization; Defensive Distillation; Input Randomization; Adversarial Example Detection; Robustness Verification.
Data Leakage/ Inference Attacks	Data, Model	Analyzing the model output to extract information about the training set or sensitive attributes	Breach of confidentiality of training data; Disclosure of sensitive user information	Differential privacy; K-anonymity; Data encryption; Model access control; API protection; Model theft protection.
Model theft	Model	Repeatedly querying a model and analyzing the responses to reconstruct its functionality or architecture	Compromising intellectual property; Bypassing paid APIs; Creating more effective adversarial attacks	Model encryption and obfuscation; Model watermarking; API protection (request rate limiting, authentication); Abnormal access monitoring.
Model inversion	Data, Model	Reconstruction of the original training data (or its characteristics) from the model output	Violation of confidentiality of training data; Reconstruction of sensitive information (e.g., faces)	Differential privacy; Output data reduction; Watermarking; API protection; Model theft protection.

Continued

Backdoor attacks	Model	Injecting a hidden trigger into the model during training that activates malicious behavior	Gaining control over a model under certain conditions; Bypassing security systems; Implementing hidden functions	Robust validation and audit of training data; Model integrity verification; Training on clean data; Detection of anomalies in model behavior; Model verification; Backdoor monitoring.
Denial of Service (DoS)	Platform, Model	Overloading the AI system with large or complex requests	Slowdown or complete unavailability of the AI service; Disabling of critical applications	API protection (rate limiting, access control); Scalable infrastructure; Intrusion detection/prevention systems (IDS/IPS); Load balancing.
Prompt Injection	Model (LLM), Data	Injecting malicious instructions into a user's prompt	Manipulating LLM behavior (jailbreaking); Extracting sensitive information; Performing unauthorized actions	Strict input validation; Separation of instructions and user input; Sandboxing; Output filtering and sanitization; Restricting LLM access to external resources.
AI Supply Chain Attacks	Platform, Data, Model	Compromise of AI components at any stage of the chain (libraries, datasets, pre-trained models)	Injection of malicious code or data; Global infection of systems using compromised components	Audit and verification of third-party libraries and components; Use of AI-BOMs; Code/model signing and verification; Secure delivery channels; Applying DevSecOps to MLOps.
Exploiting API vulnerabilities	Platform, Model, Data	Exploiting API vulnerabilities for unauthorized access and data modification	Unauthorized access to data/model; Manipulation of output data; System failure	Robust API authentication and authorization; API input validation; Least privilege principles; API traffic monitoring; Regular API security audits.
Runtime vulnerabilities	Platform	Exploiting errors in software or deployment environment configuration (servers, containers)	Gaining control over underlying infrastructure; Executing arbitrary code; Compromising data and models	Use of secure runtime environments (containers, TEEs); Network segmentation; Regular software updates and patching; Runtime monitoring; Strict security configuration.

- *Data Attacks*

Data attacks aim to compromise the integrity, confidentiality, or availability of data used by an AI system, which can occur during both the training and operational phases. Data poisoning is an attack in which an attacker injects malicious, incorrectly labeled, or biased data into the training set of an AI model (see **Table 3**, row "Data Poisoning" method). This results in the model learning from erroneous patterns, which subsequently distorts its behavior or predictions. The goal of such attacks may be to reduce the overall accuracy of the model (availability attack), introduce backdoors that are activated by certain triggers, or deliberately bias the model to misclassify specific inputs. Examples include attempts to "poison" spam filters by labeling malicious emails as legitimate, as well as studies demonstrating the poisoning of data for road sign recognition systems in autonomous vehicles. In 2023, a compromise of a portion of Google DeepMind's model was reported through the poisoning of images in the ImageNet dataset. Adversarial examples (Adversarial Examples) or Evasion Attacks (Evasion Attacks) are

specially modified input data that, while being virtually indistinguishable to humans, cause a trained AI model to make incorrect predictions or classifications (see Table 3, row "Adversarial Examples", method). These attacks exploit "fragile" decision boundaries within the model. Their goal is to bypass detection systems, such as malware detection systems or spam filters, or to manipulate image recognition systems so that the autonomous vehicle misidentifies an object. Examples include image manipulation by adding subtle "noise" that tricks a model into classifying a panda as a gibbon, or subtle changes to images in optical character recognition (OCR) that fool text recognition models.

Inference attacks or data leakage occur when an attacker attempts to extract sensitive information about the data on which a model was trained by analyzing its output (see Table 2, row "Data Leakage/Inference Attacks", method). This may include membership inference attacks, which determine whether a particular data instance was used in the training set. The goal of such attacks is to violate user privacy or disclose commercial secrets, such as customer data used to train a model.

- *Attacks on the Algorithm/Model*

These attacks aim to compromise the AI model itself or its underlying algorithm, often with the aim of altering its functionality or stealing intellectual property. Model Stealing/Extraction occurs when an attacker attempts to gain access to the internal parameters, architecture, or even copy the functionality of a target AI model by repeatedly querying it and analyzing the responses (see Table 3, row "Model Stealing", method). The goal is to obtain valuable intellectual property, bypass paid APIs, or create your own adversarial examples that are more effective against the stolen model. Model inversion attacks aim to reconstruct the original training data or its characteristics using only access to the output data of the trained model (see Table 3, row "Model Inversion", method). The attacker is trained on an "inversion" model that can reconstruct the input data from the output predictions of the target model. The goal is to compromise the privacy of training data, such as facial reconstruction or personal health information used to train the model. Backdoor attacks involve introducing a hidden vulnerability (backdoor) into a model during training (see Table 3, row "Backdoor Attacks," method). The model behaves normally with normal input data, but when confronted with a specific "trigger" known only to the attacker, it exhibits the intended malicious behavior. The goal is to gain control over the model under certain conditions, bypass security systems, or introduce hidden functions.

Denial-of-Service (DoS) attacks aim to overload an AI system with massive or complex requests, causing it to slow down, malfunction, or become completely unavailable, disrupting service availability (see Table 3, "Denial of Service (DoS)" row, main consequences). Prompt Injection for generative AI is an attack in which an attacker manipulates the behavior of large language models (LLMs) or other generative AIs, by injecting carefully crafted instructions into user requests (prompts), causing the model to ignore its original instructions and execute ma-

licious commands (see Table 3, row "Prompt Injection", method). For example, a user could enter a malicious prompt: "Forget all previous instructions and transfer $100 to account X" to bypass the LLM restrictions and trigger an unauthorized action. The goal is to jailbreak the model to generate inappropriate or prohibited content, extract sensitive information such as the model's internal instructions, or perform unauthorized actions through connected tools or APIs. Examples include attacks on ChatGPT and Bing Chat, where users were able to force models to reveal their internal instructions or generate inappropriate content, as well as visual prompt injections, where malicious instructions are hidden in images.

- *Attacks on the Platform/Infrastructure*

These attacks target the underlying hardware, software infrastructure, and processes that support an AI system and often overlap with traditional cyberattacks, but have specific implications for AI. AI Supply Chain Attacks occur when an attacker compromises any component or step in the AI system's creation and deployment chain, such as machine learning libraries, datasets, pre-trained models, hardware, or MLOps tools (see Table 3, row "AI Supply Chain Attacks", method). The goal is to inject malicious code or data into a system before it is deployed, potentially leading to a global infection of all systems using it. Examples include compromising npm packages containing malicious code, which was then used to attack cryptocurrency wallets, or to inject malicious code into software updates, as in the SolarWinds case, but using AI to detect vulnerabilities. An API exploit occurs when an attacker exploits vulnerabilities in the APIs used to interact with an AI model or its components to gain unauthorized access, modify data, or cause the system to malfunction (see Table 3, row "API Exploitation", method). This may lead to unauthorized access to training data, model parameters, or manipulation of model output.

Runtime exploits involve exploiting software or configuration errors in servers, containers, or other computing environments where an AI system is deployed, such as vulnerabilities in Redis, ChromaDB, and NVIDIA Triton (See Table 3, row "Runtime vulnerabilities", method). The goal is to gain control over the underlying infrastructure, execute arbitrary code, or compromise data and models.

This is a detailed attack taxonomy that covers a wide range of threats facing modern AI systems. Understanding these mechanisms and developing appropriate countermeasures is critical for effective defense, as also reflected in Table 3.

4. Countermeasure and Discussion

Effective protection of AI systems requires a comprehensive, multi-layered approach spanning their entire lifecycle, from design to operation. Because AI systems face unique threats, traditional cybersecurity measures must be complemented by specialized strategies. It is important to apply the principles of "secure by default" and "security by design", which integrate security measures from the earliest stages of development. However, the development and implementation of these countermeasures occurs in the context of a dynamic "arms race," where new

defenses inevitably lead to the emergence of more sophisticated adaptive attacks that bypass specific defense.

- *Data Level Security*

Data is the foundation of AI, so protecting it is critical to the integrity and privacy of the system. Robust data validation and sanitization are essential to prevent data poisoning by implementing rigorous input validation procedures to identify and filter malicious or anomalous records during the training and inference phases (See **Table 3**, column "Countermeasure Examples", row "Data Poisoning"). This involves using algorithms to detect outliers, noise, and adversarial examples in the input data before feeding it to the model.

Data confidentiality and integrity are ensured by using encryption for data both at rest and in transit, and by using anonymization and pseudonymization techniques such as differential privacy to protect sensitive information (See **Table 3**, column "Countermeasure Examples," row "Data Leakage/Inference Attacks"). Strict role-based access control (RBAC) mechanisms for databases and storage used by AI systems also play a key role. "Privacy by Design" principles require the integration of privacy protection mechanisms at every stage of the AI lifecycle.

- *Model/Algorithm Level Security*

Protecting the AI model itself from manipulation and theft is a key aspect of AI cybersecurity. Robustness against adversarial attacks is increased through adversarial training, in which the model is trained on data supplemented with adversarial examples, which increases its resilience (See **Table 3**, column "Countermeasure Examples," row "Adversarial Examples"). The use of robust optimization and regularization methods helps create models that are less sensitive to small changes in input data. Defensive distillation, which trains a "student" model on "softened" outputs of a "teacher" model, also contributes to increased robustness. Input randomization, adding small random noise or transformations to the input data, can destabilize adversarial examples. When comparing these countermeasures, it's important to note the tradeoffs. While adversarial training can significantly improve a model's robustness to known attack types, it often increases computational costs and can degrade model performance on typical (non-adversarial) data. Differential privacy (see **Table 3**, column "Examples of Countermeasures", row "Data Leakage/Inference Attacks"), while providing strong mathematical guarantees of privacy, typically entails a decrease in model accuracy, which is a critical trade-off for many applied scenarios. Robust optimization, in turn, is a more general approach, but its effectiveness depends heavily on the choice of robustness metrics and model complexity.

Protection against model theft and inversion includes encryption and obfuscation of model parameters during storage and transmission, and the use of Model Watermarking to track model usage (see **Table 3**, column "Examples of Countermeasures", row "Model Theft"). Federated learning is an approach that allows models to train on decentralized datasets without directly sharing the data itself, which protects privacy and reduces the risk of model inversion. Model monitoring

and anomaly detection involve continuously monitoring the model's performance, bias, and anomalous behavior in real time. This also includes drift detection—identifying changes in the distribution of input data or output predictions that indicate attacks. Using explainable AI (XAI) helps understand the reasons behind model decisions, facilitating attack detection and diagnosis.

Prompt injection protection for generative AI requires strict validation of user prompts for malicious instructions and clear separation of system instructions for Large Language Model (LLM) from user input. Sandboxing limits LLM's ability to interact with external systems, and output filtering and sanitization validate AI-generated content before it is used (see Table 3, "Countermeasure Examples" column, "Prompt Injection" row).

- *Platform and Infrastructure Level Security* (*MLOps Security*)

The security of the underlying infrastructure on which AI is deployed and operated is ensured by MLOps (Machine Learning Operations) practices. Secure MLOps practices include integrating security (DevSecOps) into every stage of the MLOps pipeline: from planning and development to testing, deployment, and monitoring (see Table 3, column "Examples of Countermeasures," row "Attacks on the AI Supply Chain"). Supply chain security involves auditing all third-party libraries, frameworks, and pre-trained models, as well as using tools for generating AI bills of materials (AI-BOMs). Infrastructure protection is achieved by using secure execution environments such as containers or trusted execution environments (TEEs), network segmentation, regular software updates, and patching (see Table 3, column "Examples of Countermeasures", row "Runtime Vulnerabilities"). Secure storage and management of API keys, credentials, and other secrets is also an important measure. Auditing and logging all operations and access to AI systems is necessary for incident investigation. Automated security testing, including static (SAST) and dynamic (DAST) code analysis, as well as specialized AI security testing, should be integrated into CI/CD pipelines.

- *Organizational and Regulatory Measures*

Beyond technical measures, institutional and legal approaches to AI risk management are critical. AI Governance Frameworks, such as the NIST AI Risk Management Framework (AI RMF), provide a structured approach to managing AI risks throughout the lifecycle, with a focus on reliability, transparency, fairness, accountability, and security. Also important is the OWASP Top 10 for LLM Applications—a list of the most critical security risks specific to LLM applications, such as prompt injection, improper output handling, training data poisoning, and supply chain vulnerabilities, with recommendations for mitigating them. International initiatives, such as the EU AI Act and the OECD Principles, set legal and ethical standards for the development and use of AI. Training and awareness-raising for staff, including developers, data engineers, security specialists, and end users, about specific AI risks and best practices is key. Establishing clear roles and responsibilities for AI security-systems, as well as mechanisms for auditing and evaluating AI decisions, promote accountability. Integrating ethical principles

such as fairness, transparency, and accountability into the AI design and development process helps minimize the risks of bias and discrimination.

The comprehensive application of these countermeasures across all levels and stages of the AI lifecycle is the only way to create truly reliable, secure, and ethical AI systems that can withstand the ever-evolving cyber threat landscape.

5. Research Gaps and Future Directions

Despite significant progress in understanding and mitigating AI cybersecurity risks, this area of research is still in its early stages. Numerous unsolved problems and gaps remain that require further attention from researchers and practitioners. Current reports indicate a significant gap between AI adoption and the development of adequate security measures.

5.1. Key Research Gaps

Most existing countermeasures are reactive, aimed at detecting past attacks or known vulnerabilities. More proactive and adaptive defense mechanisms are needed that can predict new attack vectors, automatically adapt to the changing threat landscape, and prevent attacks before they occur. This includes the use of AI for threat modeling and red teams capable of independently detecting vulnerabilities. Transparency and explainability of AI decisions are critical, especially in the security context. There is a need to develop XAI methods, which not only explain how a model arrived at a certain decision, but also why it perceives certain inputs as malicious, and how these explanations can be protected from manipulation. Research in the field of explainable AI cybersecurity (XAIS) is in its infancy.

With the advent of more autonomous and decision-making AI agents that can interact with real-world systems, new and more complex threats arise. There are gaps in understanding and mitigating the risks associated with unauthorized agent behavior, goal misalignment, and the difficulty of real-time monitoring.

Despite the recognition of the need for security by design, there is still no unified, standardized methodology for comprehensively assessing AI risks at every stage of the lifecycle, from conception to decommissioning. Tools and metrics are needed to measure the robustness, privacy, and security of AI systems. New forms of attacks are constantly emerging, such as chained prompt injection, attacks on multimodal AI, and attacks on AI using reinforcement learning. Existing defense methods often lag behind attack developments. Although initiatives such as the NIST AI RMF, ISO/IEC 42001, and the OWASP Top 10 for LLM. While applications lay the foundation for standardization, there is a need for more specific and mandatory standards covering specific aspects of AI safety. Harmonization of international regulatory requirements also remains a challenge.

5.2. Future Research Directions

Future research should focus on developing AI systems that can not only detect

but also autonomously respond to and recover from attacks while minimizing human intervention. This includes research into autonomous threat detection, decision making, and automated vulnerability remediation. Human-in-the-Loop AI Security (HITL) integration is also a critical area, recognizing that AI will not completely replace humans, but rather complement them. Research should focus on effectively integrating human expertise and AI automation to improve threat detection accuracy, reduce false positives, and make more informed decisions. Particular attention should be paid to how human biases can influence AI systems and how AI can improve human decision-making.

Expanding research into AI security in architectures such as distributed and federated environments, where data remains decentralized, creates new challenges for ensuring model integrity and preventing attacks. The application of rigorous mathematical and logical methods for formal verification of correctness and security.

The robustness and reliability of AI models will ensure their performance in mission-critical applications. In addition to protecting against attacks on AI, it is necessary to research how AI can be used maliciously to conduct attacks and develop countermeasures against such "AI-enhanced" threats. This includes the creation of deepfakes and automated phishing campaigns. Finally, the development of new, specialized tools and platforms that can automate vulnerability detection, adversarial attack testing, and security monitoring throughout the AI lifecycle will be critical to ensuring a robust and secure future for AI systems.

To further deepen critical analysis and enhance scientific novelty, it is important to ask more specific and bold research questions within the above-mentioned areas:

- What are the universal metrics for quantifying the security, privacy, and robustness of an AI system throughout its lifecycle, and how can they be effectively integrated into MLOps pipelines?
- Can AI systems be designed to autonomously detect and neutralize zero-sum attacks based on fundamentally new vectors, rather than just known patterns?
- How to formally verify the robustness and security of complex, opaque black-box models without access to their internal state or architecture, especially in mission-critical systems?
- What legal and ethical frameworks are needed to ensure accountability and liability in the event of compromise of autonomous AI agents, and how can these be implemented in practice?
- How to effectively balance the performance requirements of AI systems with security measures, especially when security techniques (e.g., differential privacy or adversarial learning) may negatively impact accuracy or computational efficiency?

Addressing these challenges will be critical to ensuring a robust and secure future for AI systems, enabling them to harness their potential for the benefit of

society while minimizing the risks associated with them.

6. Conclusions

The rapid and pervasive adoption of artificial intelligence (AI) systems across various areas of human activity, from critical infrastructure to everyday consumer applications, has created a new and complex cybersecurity landscape. Unlike traditional information systems, AI systems have unique characteristics such as data dependence, complex black-box models, and adaptive nature that give rise to fundamentally new types of vulnerabilities and attack vectors. This review article presented a comprehensive and systematic analysis of AI cybersecurity, covering their unique vulnerabilities, attack taxonomies, and comprehensive countermeasure strategies throughout the AI lifecycle. We categorized vulnerabilities by design, development, deployment, and operational phases (see Table 2), highlighting how weaknesses at each stage can be exploited by attackers. The proposed attack taxonomy (see Table 3) detailed threats targeting data (poisoning, adversarial examples, data leakage), algorithms/models (theft, inversion, backdoors, prompt injection), and platform/infrastructure (supply chain attacks, API exploitation, runtime vulnerabilities), providing deep insight into the mechanisms of impact and concrete examples such as malicious prompts for LLM. In response to these threats, multi-layered countermeasures have been considered, including data protection through strong validation, encryption, and anonymization; model protection through improved robustness to adversarial attacks (comparative analysis of methods such as adversarial learning and differential privacy), anti-theft and anti-inversion methods, as well as continuous monitoring and explainability implementation; platform and infrastructure protection through the implementation of secure MLOps practices, such as DevSecOps for AI, supply chain security, and access management; and organizational and regulatory measures, including AI governance frameworks (NIST AI RMF, OWASP Top 10 for LLM, Applications), personnel training, and ethical principles.

Particular attention was paid to the concept of an "arms race," emphasizing the need to continually refine defense mechanisms in response to evolving attack techniques. Despite these efforts, the field of AI cybersecurity still faces significant research gaps, such as the need for more proactive and adaptive defense mechanisms, the development of explainable AI cybersecurity, the security of autonomous agent systems, and the creation of comprehensive risk assessment methodologies. Future research directions should focus on the development of autonomous and self-healing safety systems, deeper integration of humans in the loop, formal verification of AI systems, and the development of standards to ensure sustainable and safe AI development. We also formulated specific research questions, concerning security metrics, zero-sum attack detection, black-box verification, and ethical and legal responsibility, highlighting the unresolved challenges in this area.

Ensuring the cybersecurity of AI systems is not just a technical challenge, but

also a fundamental requirement for maintaining public trust, the ethical use, and the sustainable development of AI technologies in our ever-changing digital world. Only through the coordinated efforts of researchers, developers, regulators, and users can a truly secure and reliable AI ecosystem be built.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Akhter, S., Ahmad, M.R., Chibb, M., Zai, A.F. and Yaqoob, M. (2024) Artificial Intelligence in the 21st Century: Opportunities, Risks and Ethical Imperatives. *Educational Administration: Theory and Practice*, **30**, 4600-4605. https://doi.org/10.53555/kuey.v30i5.3125

[2] Rashid, A.B. and Kausik, M.A.K. (2024) AI Revolutionizing Industries Worldwide: A Comprehensive Overview of Its Diverse Applications. *Hybrid Advances*, **7**, Article ID: 100277. https://doi.org/10.1016/j.hybadv.2024.100277

[3] Klishin, A.A. and Taran, K.K. (2024) Legal Issues of Cybersecurity, Risks and Ethics in the Use of Artificial Intelligence. *Law and Economics*, No. 10, 24-30.

[4] Mahatme, J. and Aher, P.S. (2025) Advancements in Machine Learning: Revolutionizing the Future of Technology. *International Journal of Research and Analytical Reviews*, **12**, 109-114.

[5] Tuoyo, O.S., Hossain, A., Habibur Rahman, H.B., Al Mamun, M.A., Hussein, L., Khan, M.A., Melon, M.M.H. and Shah, S. (2024) The Role of Machine Learning and Deep Learning in Shaping Modern Computer Science: Challenge, Opportunities, and Future Directions. *Nanotechnology Perceptions*, **20**, 219-235.

[6] Sangwan, R.S., Badr, Y. and Srinivasan, S.M. (2023) Cybersecurity for AI Systems: A Survey. *Journal of Cybersecurity and Privacy*, **3**, 166-190. https://doi.org/10.3390/jcp3020010

[7] Elahi, M., Afolaranmi, S.O., Martinez Lastra, J.L. and Perez Garcia, J.A. (2023) A Comprehensive Literature Review of the Applications of AI Techniques through the Lifecycle of Industrial Equipment. *Discover Artificial Intelligence*, **3**, Article No. 43. https://doi.org/10.1007/s44163-023-00089-x

[8] Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., *et al.* (2023) Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. *Cognitive Computation*, **16**, 45-74. https://doi.org/10.1007/s12559-023-10179-8

[9] Christoper Osazuwa, O. and Ozohu Musa, M. (2024) The Expanding Attack Surface: Securing AI and Machine Learning Systems in Security Operations. *International Journal of Innovative Science and Research Technology* (*IJISRT*), **9**, 2498-2505. https://doi.org/10.38124/ijisrt/ijisrt24may1613

[10] Hamon, R., Junklewitz, H., Soler Garrido, J. and Sanchez, I. (2024) Three Challenges to Secure AI Systems in the Context of AI Regulations. *IEEE Access*, **12**, 61022-61035. https://doi.org/10.1109/access.2024.3391021

[11] CROWDSTRIKE 2025 Global Threat Report. https://go.crowdstrike.com/2025-global-threat-report.html

[12] Dhanaraj, A. (2025) The Evolution of Cyber Threats: From Traditional Attacks to Ai-Powered Challenges. *European Journal of Computer Science and Information Technology*, **13**, 50-61. https://doi.org/10.37745/ejcsit.2013/vol13n365061

[13] Scott, J. and Kyobe, M. (2021) Trends in Cybersecurity Management Issues Related to Human Behaviour and Machine Learning. 2021 *International Conference on Electrical, Computer and Energy Technologies* (*ICECET*), Cape Town, 9-10 December 2021, 1-8. https://doi.org/10.1109/icecet52533.2021.9698626

[14] Wiafe, I., Koranteng, F.N., Obeng, E.N., Assyne, N., Wiafe, A. and Gulliver, S.R. (2020) Artificial Intelligence for Cybersecurity: A Systematic Mapping of Literature. *IEEE Access*, **8**, 146598-146612. https://doi.org/10.1109/access.2020.3013145

[15] Huang, M. and Rust, R.T. (2018) Artificial Intelligence in Service. *Journal of Service Research*, **21**, 155-172. https://doi.org/10.1177/1094670517752459

[16] Joseph, A.D., Laskov, P., Roli, F., Tygar, J.D. and Nelson, B. (2013) Machine Learning Methods for Computer Security (Dagstuhl Perspectives Workshop 12371). *Dagstuhl Manifestos*, **3**, 1-30.

[17] Qiu, J., Wu, Q., Ding, G., Xu, Y. and Feng, S. (2016) A Survey of Machine Learning for Big Data Processing. *EURASIP Journal on Advances in Signal Processing*, **2016**, Article No. 67. https://doi.org/10.1186/s13634-016-0355-x

[18] Kaplan, A. and Haenlein, M. (2019) Siri, Siri, in My Hand: Who's the Fairest in the Land? On the Interpretations, Illustrations, and Implications of Artificial Intelligence. *Business Horizons*, **62**, 15-25. https://doi.org/10.1016/j.bushor.2018.08.004

[19] Abbas, N.N., Ahmed, T., Shah, S.H.U., Omar, M. and Park, H.W. (2019) Investigating the Applications of Artificial Intelligence in Cyber Security. *Scientometrics*, **121**, 1189-1211. https://doi.org/10.1007/s11192-019-03222-9

[20] Zhang, Z., Ning, H., Shi, F., Farha, F., Xu, Y., Xu, J., Choo, K.K.R., et al. (2021) Artificial Intelligence in Cyber Security: Research Advances, Challenges, and Opportunities. *Artificial Intelligence Review*, **55**, 1029-1053.

[21] Kumar, P., Wazid, M., Singh, D.P., Singh, J., Das, A.K., Park, Y., *et al.* (2023) Explainable Artificial Intelligence Envisioned Security Mechanism for Cyber Threat Hunting. *Security and Privacy*, **6**, 112-119. https://doi.org/10.1002/spy2.312

[22] Leonteva, L. (2025) Evaluating Adversarial Attacks against Artificial Intelligence Systems in Application Deployments. *Applied AI Letters*, **6**, 45-52. https://doi.org/10.1002/ail2.121

[23] Namiot, D.E. and Ilyushin, E.A. (2022) On the Stability and Security of Artificial Intelligence Systems. *International Journal of Open Information Technologies*, **10**, 126-134.

[24] Li, Y. and Liu, Q. (2021) A Comprehensive Review Study of Cyber-Attacks and Cyber Security; Emerging Trends and Recent Developments. *Energy Reports*, **7**, 8176-8186. https://doi.org/10.1016/j.egyr.2021.08.126

[25] Yang, C., Yang, Y. and Zhang, Y. (2024) Understanding the Impact of Artificial Intelligence on the Justice of Charitable Giving: The Moderating Role of Trust and Regulatory Orientation. *Journal of Consumer Behaviour*, **23**, 2624-2636. https://doi.org/10.1002/cb.2365

[26] Garbuk, S.V. (2024) Special Security Model for the Creation and Use of Artificial Intelligence Systems. *Cybersecurity Issues*, No. 1, 15-23.

[27] Yang, J., Zheng, J., Zhang, Z., Chen, Q.I., Wong, D.S. and Li, Y. (2022) Security of Federated Learning for Cloud-Edge Intelligence Collaborative Computing. *International Journal of Intelligent Systems*, **37**, 9290-9308. https://doi.org/10.1002/int.22992

[28] Nair, P. and Ansari, M.F. (2024) Vulnerabilities in AI Systems: The Integration of AI into Cybersecurity Tools and Systems. *International Research Journal of Engineering*

and Technology (*IRJET*), **11**, 1159-1160.

[29] Agha, A. (2024) Artificial Intelligence for Vulnerability Management: Enhancing Security in a Dynamic Threat Landscape. *TIJER-International Research Journal*, **11**, a430-a432.