BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **25**, No 3
Sofia • 2025 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2025-0028

Cybersecurity and Artificial Intelligence: Triad-Based Analysis and Attacks Review

Olena Veprytska[1], *Vyacheslav Kharchenko*[1], *Oleg Illiashenko*[1,2]

[1]*Department of Computer Systems, Networks and Cybersecurity, National Aerospace University "KhAI", 17, Chkalov Str., 61070 Kharkiv, Ukraine*
[2]*School of Built Environment, Engineering and Computing, Leeds Beckett University, Leeds, LS6 3QS, United Kingdom*
E-mails: o.veprytska@csn.khai.edu v.kharchenko@csn.khai.edu o.illiashenko@khai.edu o.illiashenko@leedsbeckett.ac.uk

Abstract**:** *This study aims to expand the understanding of Artificial Intelligence* (*AI*) *attack scenarios and develop effective protection mechanisms against them. The triadic principle was used to investigate attacks on traditional systems and AI systems, enhance these attacks using AI, and employ AI for cybersecurity defence. By systematically analysing the interactions between these elements, we create a comprehensive set of attack scenarios and corresponding defensive strategies. Current analysis reveals distinct attack patterns and vulnerabilities associated with traditional and AI-based systems. Effective defence mechanisms and strategies were identified and tailored to various attack scenarios, leveraging AI's capabilities for improved security measures. The findings provide a structured approach to understanding and mitigating AI-related threats in cybersecurity. By mapping out the roles of AI in both attack and defence, this study offers valuable insights for developing advanced tools and methods to assess system security and enhance countermeasures.*

Keywords**:** *Cybersecurity, Artificial Intelligence* (*AI*)*, Machine Learning* (*ML*)*, Cyberattacks, Countermeasures.*

1. Introduction and related works

1.1. Motivation

Artificial Intelligence (AI) is rapidly evolving, integrating into an increasing number of products, services, and processes and becoming an indispensable tool for solving complex problems and enhancing efficiency across all areas of society and industry. On the one hand, this leads to greater automation of processes, making life easier. Still, on the other hand, modern scientific and technological progress, particularly in AI, could potentially lead to certain kinds of disasters. The negative consequences of the rapid development of AI could be irreversible and have an adverse impact on society and human life.

Firstly, Artificial Intelligence Systems (AIS) are increasingly being integrated into critical components of society, creating a new surface for attacks on the AI technology itself. Secondly, the lawfulness of AI usage raises concerns. Lawfulness refers to an AI system's ability to comply with applicable laws and regulations. Any technology can be used both legally and maliciously. Recently, there has been a rapid increase in the use of AI to enhance existing cyberattacks and create entirely new services utilising AI that aim to violate legal norms. Additionally, AI-as-a-Service is becoming more widespread, lowering the barrier to market entry by reducing the skills and technical expertise required to use AI. The Bletchley Declaration [1], signed by representatives of twenty-eight governments at the AI safety summit, emphasises the importance of regulation and ethics in AI development. This declaration unites countries for joint research and the development of new rules governing the use of AI.

To understand the full range of possibilities and limitations of AI, it is essential to examine it from multiple perspectives. This includes considering it a vulnerable system that can be attacked, as a tool for creating and enhancing cyberattacks, and as a means of defence against traditional cyberattacks [2]. This comprehensive understanding can pave the way for leveraging AI's potential to strengthen cybersecurity, offering hope in the battle against cyber threats.

1.2. Related works

During the review process, 63 studies focusing on cyberattacks and the use of AI for defence against attacks were considered, as shown in Table 1. Most reviews focus on specific attacks or approaches to ensuring cybersecurity.

Table 1. Related works analysis

Reference	Type of attack	Description
[3-11]	DDoS	The sources examine the principles of DDoS attacks, potential defence mechanisms, and both traditional methods and AI-powered solutions
[12-15]	DGA generation	The sources examine various Domain Generation Algorithms (DGA) and potential defence mechanisms based on ML
[16-19]	Fake news generation	The sources examine various types of fake news, including methods for manually detecting them and utilising automated tools, as well as the challenges involved in the detection process
[20-25]	Fake images generation	The sources examine methods for detecting fake, manipulated, or misused images using traditional and AI-powered techniques
[26-29]	Deep fake generation	The sources examine methods for detecting fake videos using AI-powered tools
[30-41]	Phishing attack	The sources examine various phases of the phishing attack lifecycle, their taxonomy, and countermeasures using multiple techniques, including both static methods and AI-powered tools
[42-51]	Poisoning attacks	The sources examine various poisoning attacks (Targeted, Indiscriminate, Backdoor) on different learning and models, such as Machine Learning (ML), Deep Learning (DL), and federated learning. They review and analyse countermeasures to mitigate poisoning attacks, focusing on limitations and complexities involved in defence strategies
[48, 51-57]	Model component stealing	The sources examine various types of privacy attacks on AI models, categorization, goals, and methods of applying attacks to obtain confidential data about the internal components of models, training data, or copies of their behaviour
[51, 58-65]	Evasion (Adversarial & Sponge) Attack	The sources examine methods of conducting adversarial attacks on ML and DL models, as well as countermeasures to mitigate the consequences of attacks or avoid them altogether

While numerous cyberattack types exist, this study deliberately focuses on a selected subset that is both highly representative and directly relevant to the role of AI in cybersecurity. The chosen attack types – DDoS, DGA, fake/disinformation, phishing, poisoning, model theft, and adversarial (evasion) attacks – were identified from the analysis of 63 recent studies (see Table 1) as among the most prevalent and impactful. They represent diverse mechanisms (network-level, content-level, data-level, and model-level), ensuring that the taxonomy captures the main dimensions of AI-related cyber threats without diluting the analysis across less significant or rarely documented attacks. This scope allows for a balanced treatment of depth and breadth while maintaining relevance to the triadic principle.

1.3. Objectives, structure, and scenario-based approach

The primary objective of this investigation is to develop an expanded set of attack scenarios and corresponding defensive strategies [2] using the proposed triadic principle, as presented in Table 2. This principle integrates three dimensions: the type of attack, the type of protection, and the type of system. All this three dimensions could be either traditional or AI-powered. By systematically combining these elements, the study enumerates all logically possible configurations of cyberattacks and countermeasures in both traditional and AI-driven contexts. The resulting taxonomy contributes to a deeper understanding of the diverse roles of AI in cybersecurity. It provides a foundation for creating assessment tools and methodologies to evaluate system resilience and the effectiveness of countermeasures.

Table 2. Scenarios

Scenario No	Scenario
1	TA – TP – TS
2	TA – AIP – TS
3	AIA – TP – TS
4	AIA – AIP – TS
5	TA – TP – AIS
6	TA – AIP – AIS
7	AIA – TP – AIS
8	AIA – AIP – AIS

Scenarios Abbreviation: T – Traditional, AI – AI-powered, S – System, P – Protection/countermeasures, A – Attack.

It should be noted that Table 2 represents the complete set of possible scenarios that emerge when applying the triadic principle. By combining the three dimensions – type of attack, kind of protection, and type of – the resulting eight scenarios form an exhaustive taxonomy. This clarification is essential, as it ensures that the analysis covers all logically possible configurations of attack and defence in the context of both traditional and AI-driven cybersecurity.

Based on a set of attack scenario taxonomy [2], the proposed methodology can be dedicated to developing a framework for cybersecurity analysis of AI-based and AI-protected systems in conditions of AI-powered attacks. To better understand the complexities of cybersecurity threats in AI-powered systems, a comprehensive analysis identified seven general attacks, along with subdistributions that cover the abovementioned scenarios. The survey organisation is presented in Fig. 1.

The first part motivates by emphasising the critical need for advanced cybersecurity frameworks tailored to AI-powered systems in response to the evolving cyber threat landscape. The second part reviews prior research efforts, primarily concentrating on categorising cyber threats and devising defence mechanisms to counter them. The third stage delineates the primary objective of proposing a methodology to analyse and combat AI-driven cyber threats. It also provides a structured overview of the research methodology, including the definition of key terms (described below), aiming to offer a comprehensive approach to addressing cybersecurity challenges in the context of AI advancements.

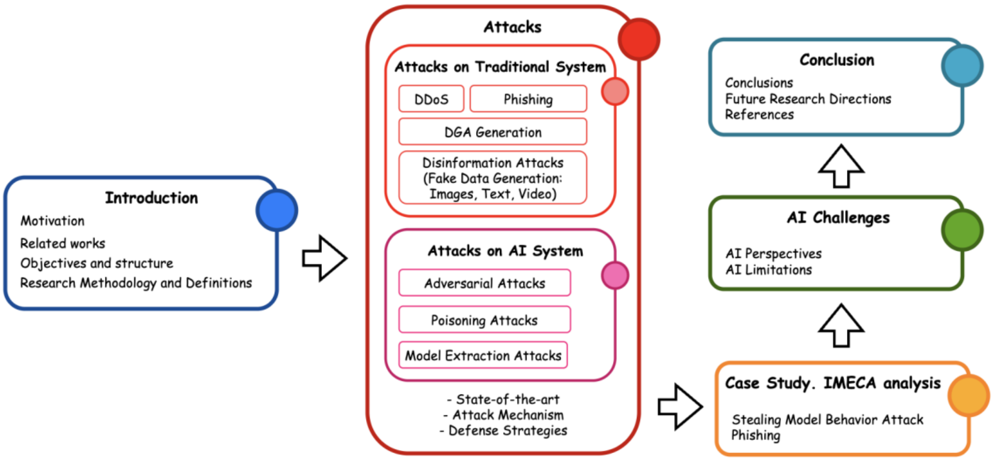


Fig. 1. Survey process and overall structure of the manuscript

In the paper, a detailed analysis is conducted for each of the defined attacks – DDoS, DGA generation, fake data generation, phishing, penetration attacks, model extraction, and poisoning attacks – to cover various crucial aspects. This includes examining state-of-the-art techniques employed in perpetrating the attack, understanding its impact on the fundamental principles of cybersecurity – namely, Confidentiality, Integrity, and Availability (CIA) – explaining the underlying attack principles, and assessing the potential for leveraging AI to enhance the effectiveness of the attack. Furthermore, defensive strategies are explored, both with and without the integration of AI, to counteract the identified threats.

The remainder of the manuscript is organised as follows. Sections 2-8 examine selected attack types, including DDoS, DGA, fake/disinformation, phishing, poisoning, model theft, and adversarial attacks. For each attack, the state-of-the-art attack mechanism and defensive strategies (both traditional and AI-powered) are analysed, along with their impact on the CIA triad. Section 9 presents a case study of risk assessment using the quantitative methodology Intrusion Modes Effects Criticality Analysis (IMECA) [2]. Section 10 discusses the broader challenges and limitations of AI in critical systems, with a particular focus on explainability, lawfulness, and ethics. It outlines the role of AI within the cyber kill chain, emphasising its importance in enhancing overall cybersecurity defence strategies. Finally, Section 11 concludes the paper and outlines future research in the field of AI-powered cybersecurity.

1.4. Definitions

To create a taxonomy, it is necessary to define what constitutes traditional and AI-powered attacks, including AI-powered countermeasures, and to distinguish between traditional and AI Systems.

Traditional System – a set of resources organised for the collection, processing, maintenance, use, sharing, dissemination, or hosting of information, without embedded capabilities of AI for analysis and decision-making.

Artificial Intelligence System – a complex of tools and algorithms aimed at collecting, processing, and understanding information, and capable of analysing data and making intelligent decisions using methods such as ML, neural networks, and DL, to achieve tasks previously considered achievable only by humans.

Traditional Cyber Attack – intentional exploitation of vulnerabilities in computer systems and networks to cause harm, gain unauthorised access, destroy information, etc.

AI-powered Cyber Attack – attacks using AI technologies to enhance existing cyber-attacks and create entirely new services using AI, aimed at violating legal norms.

Attack on AI – deliberate manipulation of AIS with the ultimate goal of causing its malfunction.

Traditional Protection – a set of methods and tools designed to protect computer systems and data from threats and attacks. TP includes the use of antivirus software, firewalls, security monitoring systems, authentication and authorisation methods, data encryption, and other traditional protection methods to ensure the confidentiality, integrity, and availability of information.

Protection based on AI for Traditional Systems – involves using AIS to detect and prevent security threats in real-time, such as detecting unusual activity in computer networks, filtering spam and malicious emails, identifying manipulated images, etc.

Protection of AI Systems using AI Tools – entails developing and implementing intelligent methods and algorithms that leverage the process of creating AI models to detect and prevent potential threats for self-defence.

2. DDOS attacks

2.1. State-of-the-art

Research into Distributed Denial of Service (DDoS) attacks remains a popular and active field of study: [3] describes the lifecycle, taxonomy, architecture, and characteristics of stationary and mobile botnets, and it outlines requirements for defence mechanisms against DDoS attacks and compares ML and statistical methods; [4] analyses Botnet Detection Techniques, botnet Command and Control

(C&C) system architecture, and more; [8] presents an overview of the latest methods for detecting DDoS attacks at the application level; authors in [9] delve into performance evaluation metrics, attack execution tools, and systematic defence mechanisms against DDoS.

2.2. Attack mechanism

DDoS attacks are typically launched using botnet technology through centralised, distributed, or hybrid command-and-control architectures [4]. The botmaster can maintain a Command and Control (C&C) server to manage the bot and initiate various types of cyberattacks, such as DDoS attacks, spam, phishing, fraud, information theft, and cryptocurrency mining. The operation scheme of a DDoS attack is illustrated in Fig. 2. The architecture of a botnet's C&C system is generally categorised into three types [4]: centralised – where the bot primarily receives control commands from a polling-based control server, and the botmaster sends control commands and resources to the zombie hosts through these servers; distributed – where any node can act as both a client and server simultaneously and the communication process doesn't rely on publicly reachable server resources, hybrid – which typically combines both central and Peer-to-Peer (P2P) structures. Attackers leverage large-scale compromised hosts, known as "zombies", to flood a target with traffic, rendering services unavailable.

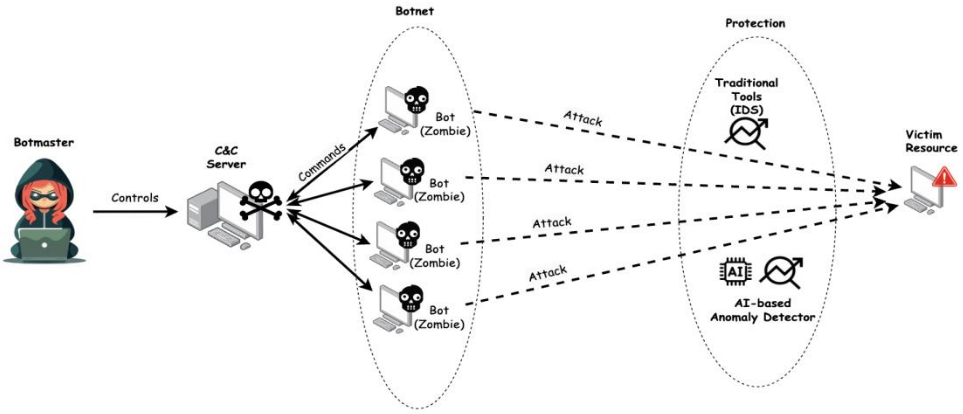


Fig. 2. DDoS attacks scheme

2.3. Defence strategies

TTraditional countermeasures include IDS, firewalls, and anomaly detection; however, these approaches struggle to distinguish between malicious and legitimate traffic. AI-based methods – such as ML classifiers (Naive Bayes, Random Forest, SVM) and DL models (CNNs, RNNs, LSTMs) – improve detection by learning complex traffic patterns and achieving lower false positive rates [9-11]. This shift from traditional filtering towards AI-driven solutions represents the core contribution of current research on DDoS mitigation.

In summary, although DDoS is a well-studied attack, AI substantially increases both the scale and efficiency of botnet-based assaults while also enabling more accurate and adaptive detection. AI-driven models, particularly ML and DL approaches, outperform traditional defences in distinguishing malicious from legitimate traffic, making AI the central factor in both the evolution and mitigation of modern DDoS threats.

3. DGA generation attack

3.1. State-of-the-art

Domain Generation Algorithms (DGAs) are utilised by various malicious software families to periodically generate new domain names for connecting infected hosts to their C&C servers. DGAs may employ pseudo-random algorithms to create domain names, starting from a common root/initial seed, which could be a string, a date, or a set of numbers and special characters. DGAs are integral to DDoS attacks.

In the context of DGA research, the primary focus is on methods for distinguishing between legitimate and malicious domains. Article [13] describes techniques for detecting generated domain names, particularly those used by malicious software. Article [12] discusses the taxonomy of DGAs, including Script-Based DGAs and Binary-Based DGAs. Article [14] proposes a method for detecting DGA domain names based on ML.

This rotation of domains complicates blacklisting and enhances botnet resilience [12-15]. As this attack vector is extensively analysed in the literature, we provide only a brief overview before focusing on the specific role of AI in both creating and defending against DGA-based threats.

3.2. Attack mechanism

DGAs generate thousands of domain names daily using pseudo-random or algorithmic seeds shared between the malware and its C&C server. Attackers register a fraction of these domains, ensuring persistence even if some are blacklisted. The operational principle of DGA is illustrated in Fig. 3.

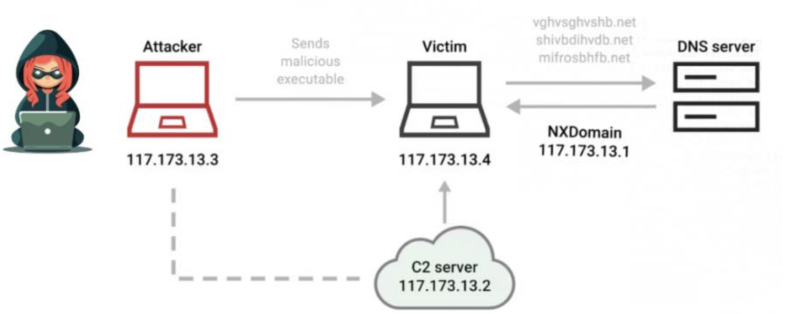


Fig. 3. DGA attack scheme

In addition to traditional domain generation algorithms, their generation can be automated using AI. Recent advances include AI-powered DGAs, such as

DeepDGA, which leverage GANs to produce domain names that closely resemble legitimate ones and evade detection even by DL-based classifiers [66].

In summary, DGAs remain a key enabler of resilient botnet infrastructures, and the integration of AI significantly increases their sophistication by generating domains that closely resemble legitimate ones and evade traditional detection. At the same time, AI-powered defence mechanisms – ranging from feature-based ML classifiers to deep learning models – provide the most effective countermeasures. This dual role underscores the importance of advancing adaptive AI-driven detection strategies to keep pace with the evolution of AI-enhanced DGAs.

3.3. Defence strategies

Traditional methods, such as blacklisting or rule-based filtering, are ineffective against rapidly changing domains. Instead, detection has shifted towards AI approaches. Feature-based ML models use lexical and statistical properties (length, character ratios, etc.) [13, 14], while featureless DL methods employ character-level embeddings for classification [15]. These AI-driven approaches significantly outperform traditional defences in terms of adaptability and detection accuracy, making them the primary direction for combating modern DGAs.

4. Disinformation attacks

4.1. State-of-the-art

The EU Commission [67] defines misinformation as the creation, presentation, and dissemination of false or misleading information for economic gain or intentional public deception, which can harm society. Disinformation, encompassing fake news, manipulated media, and synthetic content, represents a critical cybersecurity and societal challenge. It can undermine public trust, influence political processes, and facilitate large-scale manipulation campaigns. While numerous typologies exist (misinformation, propaganda, satire, hoaxes, rumours, clickbait) [16-19], this study provides only a concise overview and instead emphasises the role of AI in generating and countering disinformation.

4.2. Attack mechanism

Social bots are the most common non-human creators of fake news. Social bots are computer algorithms designed to mimic human behaviour, automatically generate content, and interact with people on social media. Creating fake news involves various AI techniques, such as Generative Adversarial Networks (GAN), models like Generative Pre-trained Transformer (GPT), Long Short-Term Memory (LSTM), and Natural Language Processing (NLP) techniques to generate false information. Real people are vital sources of spreading fake news; they may use AI to create and spread fake news, but they can also rely on other methods. Disinformation can take various forms, including textual data, images, videos, and audio recordings, which will be discussed in the following sections.

4.3. Defence strategies

When analysing the news, key characteristics can be identified [16, 17]: Source (human or automated), information content, social context, target promoters (bots or humans), and target audience or victims [18].

4.3.1. Detection of fake news

Fact-checking determines the accuracy and truthfulness of information that may be unreliable or contain falsehoods. Fact-checking can be conducted manually and automatically [19], utilising computer algorithms and AI to analyse and verify information, detecting false or unreliable facts in news and other sources. Online news can be collected from various sources, such as news agencies' homepages, search engines, and social media websites [16, 18]. Additionally, models are aimed at detecting toxic content, manipulation, hate speech, sexism, anti-Semitism, radical symbols, and fake reviews.

4.3.2. Fake image detection

Fake or manipulated images are also used to create false news. The complex changes may involve removing or inserting people into images [20]. Visual features include clarity assessment, coherence assessment, similarity distribution histogram, diversity assessment, clustering assessment, and more. In [17], it is noted that compared to genuine images, fake images often exhibit higher clarity and coherence but lower diversity and clustering.

Traditional methods for detecting fake images (or those used in the wrong context) include Metadata analysis [21], which includes information about how files are generated and processed, and image reverse search based on the web.

AI methods for detecting fake images (or those used in the wrong context) are as follows.

- **Image forgery detection [22].** The goal is to detect image manipulations, including copying and pasting, montage, deletion, and enhancement. Authors of [22] propose a forgery detection system implemented through three modules: A metadata-based module, neural networks, and error-level analysis. Meanwhile, resource [23] discusses the creation of the Digital Forensics 2023 (DF2023) database, which shall consist of one million images distributed across four main categories of forgeries, thereby opening avenues for research and the development of new forgery detection methods.
- **Image-based geolocation estimation.** The task involves estimating the location where a photo was taken. Methods for this include location estimation using image content and DL to recognise patterns.
- **Fake face detection.** The task is to detect counterfeit images of faces, which can be used to create fake user profiles, author photos, or those of witnesses or experts.

GANs can generate realistic fake face images that can easily deceive people. For example, the service **https://thispersondoesnotexist.com** creates realistic photos of people who do not exist using DL technology.

Research [24] focuses on understanding and detecting fake faces created by GANs, making key contributions: first, studying how humans and Convolutional Neural Networks (CNNs) distinguish between fake and real faces; second, enhancing fake face detection through the “Gram-Net” architecture, which utilises global texture representations. [25] discusses a new method to thereby enhance fake face detection and enable more effective detection of counterfeit details.

4.3.3. Fake video detection

Deepfake is a technique for synthesising human images based on AI. Deepfake combines and overlays existing pictures and videos onto target images or videos using ML methods. These realistically manipulated videos are difficult to distinguish with the naked eye [26].

However, deepfakes also pose significant threats to our society, political systems, and businesses According to the purposes of facial manipulation algorithms, existing deepfake algorithms can be categorised into two categories [27]: Face swapping and face reenactment.

The goal of AI-based methods is to detect fake or altered video content, including identifying manipulated video recordings and content created within videos. This requires various technologies, such as pattern analysis and recognising fake elements or artefacts that may indicate video manipulation. In [28], a proposed system employs a neural network approach to detect Deepfake videos. In [29], an automatic deepfake detection system is introduced, utilising a CNN to extract frame-level features. These features are then fed into a Recurrent Neural Network (RNN), which determines whether the video has been manipulated.

In summary, while disinformation is not a new phenomenon, AI dramatically increases its scale and realism. At the same time, AI provides the most promising defensive capabilities, making the balance between adversarial generation and detection a dynamic research frontier.

5. Phishing attack

5.1. State-of-the-art

Phishing is one of the most widespread cyberattacks, exploiting both technical and human vulnerabilities to obtain sensitive data, spread malware, or hijack accounts.[68]. Current research on phishing attacks and defences includes the development of various taxonomies, classifications, and attack stages [30-33]. Relevant directions involve developing and assessing countermeasures based on ML, DL, and traditional methods [34-40]. The emergence of Phishing-as-a-Service (PhaaS) platforms further lowers the entry barrier for adversaries by offering ready-to-use infrastructures [30].

5.2. Attack mechanism

Among the methods of phishing attacks, the following are distinguished: deceptive phishing, where the perpetrator uses social engineering methods to deceive victims [31-32]. The most common types of social phishing include phishing emails, spoofed websites, phone phishing (vishing and SMishing), social media attacks (soshing and social media phishing), and technical phishing.

It is possible to highlight the diversity of phishing attacks via email, which consists of six stages, as shown in Fig. 4, as it is the most widespread variant.

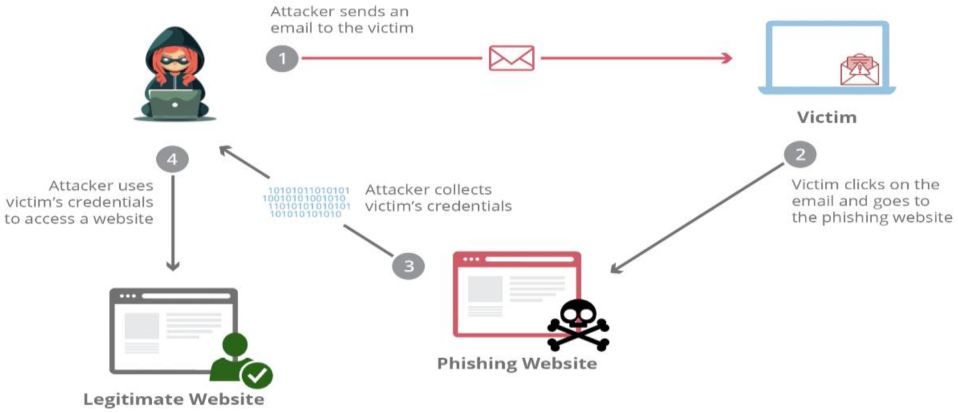

Attacker sends an email to the victim
1
Victim
2
Victim clicks on the email and goes to the phishing website
3
Attacker collects victim's credentials
4
Attacker uses victim's credentials to access a website
Phishing Website
Legitimate Website

Fig. 4. Phishing attacks scheme

Regarding the creation of phishing text, two actors are identified: a human or a bot that manually alters text using computer programs or generates it automatically using AI. AI significantly extends these capabilities: frameworks like SNAP_R [41] automatically generate spear-phishing emails tailored to specific users, while advanced natural language models create highly convincing phishing messages. Automated phishing generators, originally developed for defensive dataset creation [36, 39], can also be misused for malicious purposes. The combination of PhaaS platforms and AI text generators increases the scalability and success rate of phishing campaigns.

5.3. Defence strategies

5.3.1. Traditional methods of phishing detection

There are traditional computerised anti-phishing techniques and AI-based methods of non-computerised and computerised anti-phishing techniques [30, 34], yet these are insufficient against AI-enhanced phishing. The primary focus in phishing detection is classifying objects, such as websites, emails, links, and messages, to determine whether they are phishing or legitimate. ML trains models that learn to distinguish phishing from unknown or legitimate objects [35].

Current defences increasingly rely on AI-based techniques:

- Machine learning models (logistic regression, SVM, random forest, gradient boosting) classify emails, websites, and URLs using features such as HTML content, hyperlinks, or lexical cues [35, 36, 39].
- Deep learning approaches (CNNs, LSTMs, hybrid ensembles) enable detection of previously unseen phishing samples and adapt to new attack patterns [37, 40].

- Generative approaches (e.g., Adversarial Autoencoders) are employed to augment training datasets with realistic phishing samples, improving resilience against novel attacks [38].
- Explainable AI methods improve trust and usability by highlighting which features lead to classification decisions.

5.3.2. Attacks on AI

The success of AI in a wide range of applications heavily relies on the availability of computational resources and large datasets for training. However, the fundamental assumption that "training datasets are representative and adequately reflect the conditions encountered during real-world testing of AI systems, and the basic parameters and structure of AI models are considered secure by default", upon which training is based, creates vulnerabilities. However, this assumption is invalid in data poisoning attacks, model component theft, or adversarial attacks. The main challenges include the complexity and cost of data collection and labelling, as well as the setup and deployment of custom AI models. In such cases, these tasks are often delegated to third parties and services, which can lead to potentially negative and risky consequences.

AI system vulnerabilities stem from: (1) using unreliable external data sources [42, 45]; (2) reliance on third-party training platforms that attackers can alter during execution [45]; (3) retraining with feedback, which enables poisoning through false or biased inputs; (4) employing third-party pre-trained models, where infected networks may be distributed via APIs, limiting defender control [49]; (5) manipulation of local models or parameters in federated learning, allowing adversaries to compromise clients or upload malicious models [47]; and (6) model memorisation of training data, enabling reconstruction of sensitive information [51].

Attacks on AI are becoming increasingly common and must be factored into risk assessments. They can be categorised as [42]: (1) integrity violations – evasion without disrupting normal operations; (2) availability violations – denial of service by impairing functionality; (3) confidentiality violations – extraction of private system or user data.

In summary, while phishing has long been a significant cyber threat, the introduction of AI both intensifies the threat (through automation and personalisation) and provides the most promising countermeasures for detection and prevention.

6. Poisoning attacks

6.1. State-of-the-art

Poisoning attacks manipulate training data or the training process to degrade model performance. They are commonly classified as indiscriminate, targeted, and backdoor attacks [42]. Studies propose taxonomies of attacks, defence methods, and analyse transferability issues. For example, [43] compares the impact of poisoning across multiple ML and NN models.

Surveys [44, 45] address backdoor attacks, their creation, detection, and defence, with [45] comparing them to adversarial and data poisoning, while [44] reviews attacks across the ML pipeline. Federated learning is also vulnerable, as shown in [46-48], which details attack mechanisms and possible defences. A conditional taxonomy of poisoning attacks based on various classification criteria is provided in Fig. 5 [42, 43].

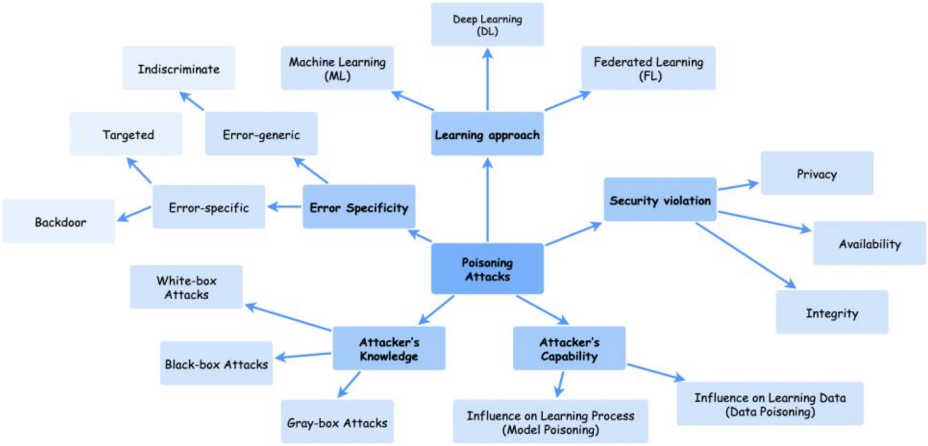


Fig. 5. Taxonomy of poisoning-based attacks with different categorisation criteria

6.2. Attack mechanism

6.2.1. Indiscriminate poisoning attacks

In indiscriminate poisoning attacks, the attacker's goal is to misclassify benign test samples by adding new malicious samples or modifying existing ones in the training dataset, as shown in Fig. 6a. Existing methods are based on different assumptions about the attacker's knowledge and capabilities [49]: (1) the attacker has access to both training and test data, the target model, and the training procedure; (2) the attacker has access only to the training data, the target model, and the training procedure; (3) the attacker has access only to the training data.

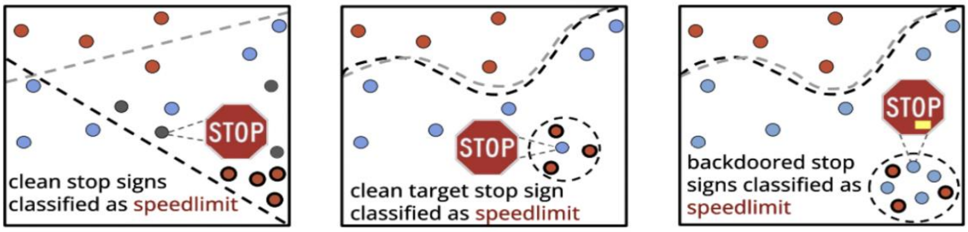
One of the key factors for the success of a data poisoning attack is the ratio of malicious participants (for FL) and the volume of data they corrupt, which applies to both ML and federated learning. According to [48], the success of a poisoning attack increases linearly with the number of corrupted samples (both FL and ML). [49] analyses the efficacy of adding a few corrupted data points to the primary training dataset, although some works consider the possibility of modifying the entire primary dataset. Currently, examples of indiscriminate poisoning include Label-flip Poisoning, Bilevel Poisoning, Bilevel Poisoning (Clean-label), and Dirty Label Attack [42, 46].

6.2.2. Targeted poisoning attacks

In targeted poisoning attacks, the attacker again manipulates a subset of the training data, but this time to misclassify a specific set of test samples, as shown in Fig. 6b. The difference from indiscriminate poisoning is that targeted attacks aim to preserve the availability, functionality, and behaviour of the system for legitimate users while causing misclassification of specific target samples. These attacks do not require changes to the test data but manipulate the training data. The review [42] discusses various attack strategies: (1) Bilevel Poisoning, that involves the attacker creating targeted attacks aimed at specific goals by manipulating the training data. The main idea is that the attack is optimised on a set of validation samples and evaluated on a separate set of test samples; (2) Feature Collision (Clean-label), thatemploys a heuristic approach aimed at influencing models by poisoning the training data. The limitation of this approach is the requirement for a precise understanding of the feature function and its stability during training.

6.2.3. Backdoor attack

During a backdoor attack, the training data is manipulated by adding poisoned samples that contain a specific pattern, known as a backdoor trigger, which is labelled with a class selected by the attacker. This is illustrated in Fig. 6c. This typically causes the model to learn a strong correlation between the backdoor trigger and the class label selected by the attacker.


STOP
clean stop signs classified as speedlimit
STOP
clean target stop sign classified as speedlimit
STOP
backdoored stop signs classified as speedlimit

(a) Indiscriminate attack.

(b) Targeted attack.

(c) Backdoor attack.

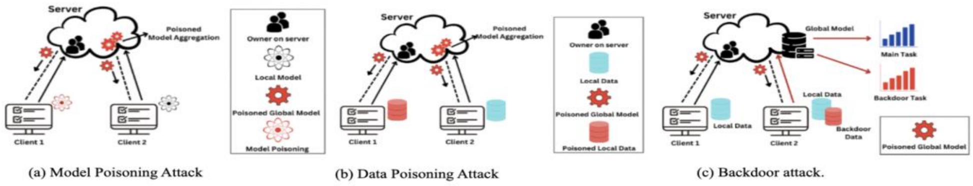
Fig. 6. Scheme of poisoning attack in machine learning [42]

Studies examine multiple methods for creating backdoor attacks. Key concepts include benign and infected models, triggers, and attack samples [45]. Backdoor poisoning is based on embedding triggers into poisoned data, with subtypes including Invisible, Static, Dynamic, and Semantic Attacks [44, 45].

The main challenge is optimising adversarial data. Techniques include multi-level optimisation to maximise trigger impact and generating universal perturbations that affect models across datasets [45]. Effectiveness is usually measured by Clean Accuracy (CA), the accuracy of benign samples, and Attack Success Rate (ASR), the proportion of poisoned samples classified as the attacker's target.

Attacks on model poisoning, as illustrated in Fig. 7a, involve directly influencing the model updates that clients send to the server. These attacks can employ various methods, such as random weight generation, Optimisation methods, or Information leakage, to modify these weights. The backdoor attack, as illustrated in Fig. 7c, aims primarily to embed a secondary task within the model. In other words, a successful backdoor attack ensures the performance of the primary task while adding a secondary task to it.

Data poisoning in any learning faces two key challenges: High computational complexity of optimisation attacks and unrealistic threat models [42]. Some attack models assume complete control over the entire training dataset, which is unlikely in real-world scenarios.



(a) Model Poisoning Attack (b) Data Poisoning Attack (c) Backdoor attack.

Fig. 7. Scheme of poisoning attack during federated training [47]

6.3. Defence strategies

6.3.1. Training data sanitisation

Training data sanitisation aims to detect and remove poisoned samples before training the model, thereby reducing their impact on the training process [42]. The primary concept of this defence is to distinguish poisoned samples from the rest of the training data, as they exhibit anomalous behaviour in the context of the data distribution. Such defences require access to the training data, and in some cases, clean validation data, to effectively identify the poisoned samples. Examples include the creation of micromodels and RONI [50], as well as blind removal strategies for backdoors [44].

Micromodel-based Defence. This approach involves training classifiers on non-overlapping epochs of the training set (micromodels) and evaluating them on the training set. By using majority voting from the micromodels, training data instances are marked as safe or suspicious.

Reject On Negative Impact (RONI). This method measures the impact of each suspicious training data instance individually and rejects those that have a significant negative impact on the overall model performance. RONI sets a threshold by observing the average negative impact of each instance in the training set and flags an instance if its impact exceeds this threshold. This threshold determines RONI’s final effectiveness and its ability to identify poisoned samples.

6.3.2. Robust training

There are also approaches to mitigate the impact of poisoning attacks during model training.

The following defence strategies are to be highlighted [42]:

(1) Model Inspection aims to determine whether the model contains a hidden trap. It involves methods for analysing the model’s output data, detecting atypical properties, or using interpretative techniques.

(2) Model Sanitisation involves removing its impact on the model.

(3) Trigger Reconstruction consists of attempting to reconstruct the hidden trap in the model for further analysis.

In summary, poisoning attacks exploit vulnerabilities in training data and processes, threatening the integrity, availability, and confidentiality of AI systems. While attackers increasingly use advanced optimisation and adaptive methods, AI also plays a crucial role in defence. Techniques such as data sanitisation, robust training, and model inspection leverage AI to detect anomalies, resist malicious data, and maintain trustworthy learning, reinforcing the resilience of AI-based cybersecurity.

7. Model extraction theft/extraction attacks

7.1. State-of-the-art

Model Theft Technique (also known as "model extraction") aims to obtain confidential data, such as training hyperparameters, model architecture, training data, or approximation of the model's behaviour, belonging to the legitimate owner of the model. These attacks expand privacy risks and can be used for the unlawful leakage of confidential information in the field of ML.

In the case of privacy attacks, vulnerabilities are identified by the models' tendency to memorise their training data [51], making them susceptible to both passive and active inference attacks, as well as the possibility of accidental data leakage, which can lead to the successful reconstruction of other clients' data.

Studies [52-54] provide a categorisation and comparison of model extraction attacks, evaluate their effectiveness, and consider appropriate defence methods under various conditions. It is also analysed in [52] that specific defence measures lose effectiveness under current attack strategies. The authors [53] propose, in addition to a taxonomy of attacks and defences, a threat model for privacy and confidentiality attacks against machine learning systems. The study [56] proposes a methodology called ML-Based Stealing Attack (MLBSA) for stealing controlled information using ML methods based on the cyber kill chain. Besides model theft attacks on ML, privacy threats to federated learning are also considered [48].

Conditionally, privacy attacks can be categorised as shown in Fig. 8.

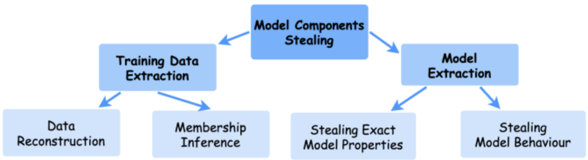


Fig. 8. Stealing/extraction attacks classification

According to [48], federated learning can be targeted by Feature inference attacks, Membership inference attacks, and Property inference attacks. These attacks aim to infer sensitive information about the training data used in federated learning scenarios.

7.2. Attack mechanism

Successful attacks focus on analysing the functionality of ML services, such as MLaaS, by giving specific input queries that exploit vulnerabilities. ML models, especially those hosted in MLaaS systems, are vulnerable to attacks because the confidentiality of ML models can be breached through pay-per-query systems in cloud services [56]. One strategy involves parameter extraction attacks using the witness identification method [52]. Another approach is to train a surrogate model, as shown in Fig. 9, using data labelled by the target model. The surrogate model is used to study the internal structure and behaviour of the target model, as well as for adversarial attacks. This approach is applicable across various types of ML models [52-54].

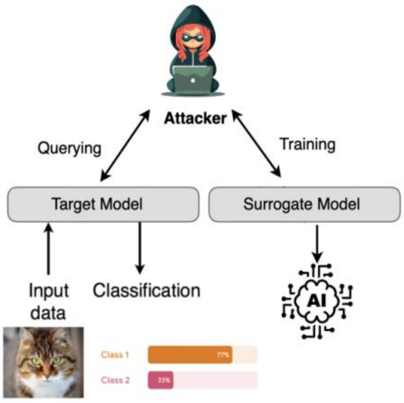


Fig. 9. Stealing/extraction attacks scheme

In addition to query-based and response analysis attacks, another mechanism considered is side-channel attacks on the privacy of ML models [52].

Access to the device’s hardware opens possibilities for more sophisticated side-channel attacks. These attacks can lead to unintended physical leaks, as illustrated in Fig. 10. The emanations manifest as physical signatures of timing reactions, power consumption, or electromagnetic emissions during data manipulation.

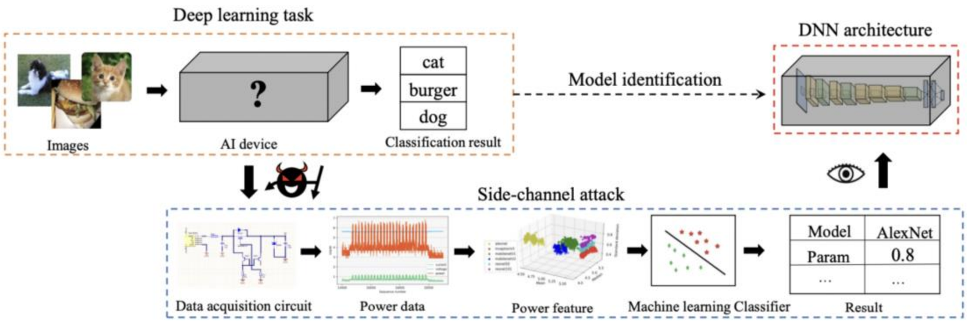


Fig. 10. Scheme of Side-channel attacks against a neural network [57]

7.3. Defence strategies

To protect models from attacks, defences can be reactive (detecting or confirming attacks) or proactive (preventing anticipated attacks). Reactive methods involve proof of ownership to establish ownership over a stolen model and attack detection to identify current threats. Reactive methods do not prevent theft itself, but rather notify the owner of the event (e.g., Watermarking, Unique Model Identifier) [52]. Proactive methods aim to prevent future attacks by modifying aspects of the model. Model extraction attacks typically require the attacker to make a series of queries to the target model. Therefore, protective measures are explored to detect these queries (Differential Privacy, Obfuscation, Prediction Vector Tampering, Regularisation). Sources also mention the use of encryption to ensure confidentiality [55] (e.g., Encrypting training data and the ML model).

WaterMarking (WM) for models involves embedding hidden information directly into the model as a method of proof of ownership. The Unique Model Identifier (UMI) is an approach used to confirm ownership of a model by defining a unique property of the model that is transmitted to a surrogate model during its theft. This approach can protect against model theft while maintaining confidentiality and ensuring data security. Differential Privacy (DP) defines a concept of confidentiality that limits how much an attacker, having access to the output of an algorithm, can learn about each record in a dataset. The original definition of DP includes a privacy parameter (budget) that restricts the probability that an attacker, with access to the algorithm's output, can determine whether a specific record was included in the dataset [51]. Methods to protect models also employ obfuscation strategies [52, 53], which involve adding perturbations to input/output/training data to complicate the process of extracting parameters from trained ML models. Encryption-based protection methods involve encrypting either training data or models, with homomorphic encryption being a key technique [55].

8. Evasion attacks (adversarial)

8.1. State-of-the-art

It is currently impossible to determine for which data or under which conditions the model will work reliably and correctly. This technological limitation forms the basis for adversarial attacks. The main goal of evasion attacks is to create adversarial examples – test samples whose classification can be altered during deployment. Additionally, deep neural networks also demonstrate vulnerability to adversarial interventions [59], which poses new challenges in the reliability and security of such models. Adversarial attacks lead to AI system malfunctions by altering the input data fed into the system, thereby compromising its performance and accuracy. Considering that decision-making systems based on AI are used in Autonomous Vehicles (AV), even small adversarial examples can cause disasters. Suppose evasion attacks are successfully used against AV. In such cases, attackers can easily manipulate data and the environment to create traffic accidents, thereby posing a threat to the personal safety of individuals [60].

Adversarial attacks can be carried out against various types of systems, including image analysis, sound, text, video, executable files, and language models, making them one of the most significant threats. Evasion attacks can be conditionally classified by criteria as shown in Fig. 11.

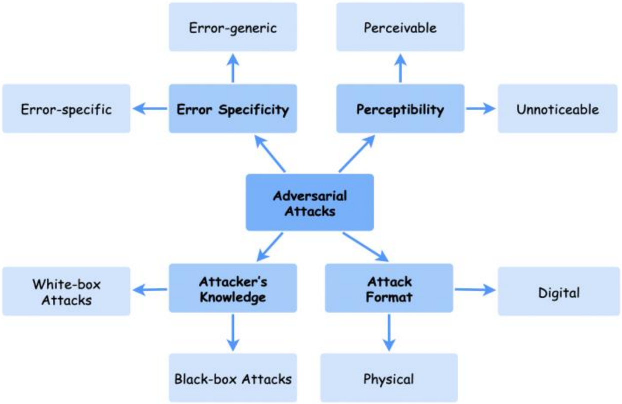


Fig. 11. Taxonomy of adversarial-based attacks with different categorisation criteria

8.2. Attack mechanism

The feasibility of implementing adversarial attacks in the physical world opens up a wide range of possibilities that can be used to mislead ML models. Scientific research has utilised and demonstrated various applications of adversarial examples in real-world scenarios [59, 62, 63]. The implementation of adversarial attacks in the digital realm requires specific algorithms that generate perturbations (digital noise).

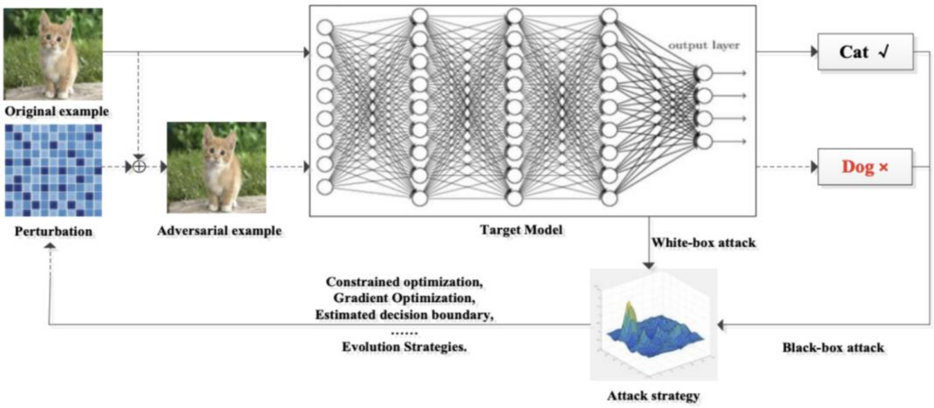


Fig. 12. Adversarial attacks scheme [62]

In modern ML and DL systems, adversarial attacks exploit the exact mechanism used for training algorithms, namely, gradient descent, to minimise the loss function. However, in the case of creating adversarial perturbations, this process is inverted. Instead of minimising the system's errors, the attacker focuses on maximising the likelihood of incorrect classification on specific data samples, as shown in Fig. 12. This is referred to as loss maximisation. Using the Projected Gradient Descent (PGD) method in the context of adversarial attacks leverages the model's gradient to determine the direction of changes that can be made to the input data to disrupt the model's predictions. The primary objective of using PGD is to identify the optimal modifications to the input data to cause the most significant disruption to the model's predictions. A perturbation budget is assumed, which defines the limit on the amount of changes that can be made. Other modern techniques/algorithms for creating adversarial examples are described in [51, 62, 64, 65].

8.3. Defence strategies

Mitigating adversarial examples remains a significant challenge, as defences that are often effective against weak attacks are later bypassed by stronger ones. Therefore, countermeasures must be tested against strong adaptive attacks, and guidelines exist for thorough evaluation. Common approaches include adversarial training, randomised smoothing, and formal verification [51]. Other methods are applied in specific cases, such as adversarial example detection, input reconstruction, gradient masking, and defensive distillation [61].

One effective method is adversarial training, a general technique that extends the training data with adversarial examples generated iteratively during the training process. The more powerful the attacks used on adversarial examples, the more robust the trained model becomes. The primary drawback of this approach is that it requires substantial resources due to the iterative generation of adversarial examples during training, thereby increasing both time and resource consumption.

Another relatively reliable method is the rejection/detection of adversarial examples. Defence systems aimed at detecting and rejecting adversarial examples rely on a thorough analysis of input samples to identify potentially harmful modifications. This can include analysing sample characteristics, such as gradients or deviations from normal ranges, or using the internal structure of the model to detect anomalies.

An expanded set of attack scenarios and defences against them for traditional IT systems (without AI) and AI-based systems, as protected assets, is provided in Appendix A and Appendix B, respectively.

9. IMECA analysis

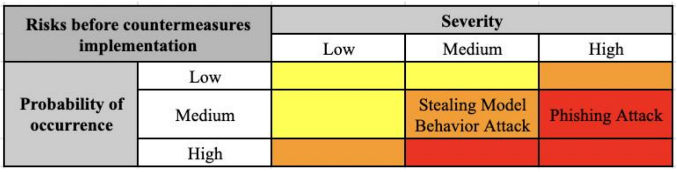
As previously highlighted, an effective framework is required for risk management and assessment to identify, evaluate, and manage risks systematically. This chapter outlines the development of an IMECA analysis [2], as shown in Table 3, and the risk matrix, as illustrated in Fig. 15, providing a structured approach to managing risks associated with phishing attacks and Stealing Model Behaviour (SMB) attacks.

By creating criticality matrices, we can assess the potential impact and likelihood of these attacks, thereby gaining a better understanding of their respective risks. The matrices highlighted that Phishing Attacks had a medium likelihood due to the proliferation of Phishing-as-a-Service platforms, automation through AI, and the use of effective social engineering techniques, all of which lower the barrier to entry and increase the efficiency of phishing campaigns, and high severity due to their potential to compromise sensitive information and cause financial losses. While Stealing Model Behaviour Attacks posed a medium likelihood and seriousness due to the extensive time, specialised knowledge, and significant resources required to execute them. Currently, there are no automated, low-cost solutions available for defending against attacks on AI. The temporary absence of concepts like "Adversarial-Attacks-as-a-Service" or "Stealing-Model Behaviour-as-a-Service" is because these types of attacks require a deep understanding of machine learning, access to specialised equipment, and substantial computational resources and time. This high level of complexity makes these methods less accessible to a broad range of malicious actors, setting them apart from more common and automated phishing attacks. The results of the cut version of IMECA analysis of cyberattacks are presented in Table 3.

Table 3. Extraction of the IMECA analysis of cyberattacks (Phishing Attack, Stealing Model Behaviour Attack) to ensure system security

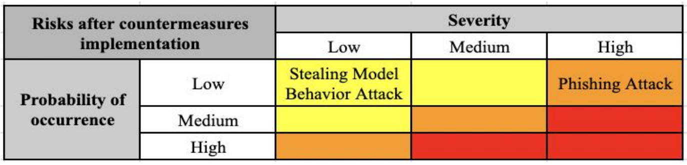
Attack	Threat	Vulnerability	Effects	Countermeasures
Phishing attack	Emails, information on social networks, phishing landing pages, pop-ups, and targeted advertising that contain a phishing link. Phishing text messages. Phone calls from spoofed numbers. Rogue mobile applications	Lack of awareness about phishing methods, absence or low effectiveness of technical security measures (e.g., anti-phishing filters), human factors, or lack of cybersecurity policies and training within the organisation	Potential effects of falling victim to a phishing attack include compromised accounts, financial losses, identity theft, malware infections, reputational damage, and legal and regulatory repercussions	Detection using AI: Utilise machine learning models for phishing classification and employ deep learning models to generate training data. Detection by traditional methods: improving user awareness about phishing, implementing proper legal protections, maintaining blacklists and whitelists, and implementing visual similarity detection for resources
Stealing model behaviour attack	Requests to the target model and collection of results for surrogate training	The propensity of models to memorise their training data and the potential for accidental data leakage lead to the successful reconstruction of other private data	Stealing model behaviour to replicate the successes of the original model or creating a surrogate to perform attacks. Using model predictions to disclose the confidentiality of sensitive records	Detection of requests and blocking. Differential privacy. Unique Model Identifier (UMI). Avoidance/obfuscation of score provisioning as part of classifier output

The analysis demonstrates that the systematic identification of attacks and the implementation of appropriate countermeasures enable a quantitative risk reduction assessment following specific mitigation strategies. After identifying these risks, we calculated the risk reduction that would result from implementing targeted countermeasures. For Phishing Attacks, countermeasures contributed to a significant decrease in likelihood by avoiding and blocking phishing sources. Similarly, Behaviour Attacks, implementing secure model training practices, monitoring for anomalous behaviour, and restricting access to sensitive model data resulted in a notable decrease in risk, as shown in Figs 13 and 14.



Risks before countermeasures implementation		Severity		
		Low	Medium	High
Probability of occurrence	Low	🟨	🟨	🟧
	Medium	🟨	🟧 Stealing Model Behavior Attack	🟥 Phishing Attack
	High	🟧	🟥	🟥

Fig. 13. Criticality matrix of cyber risks of systems before implementation of countermeasures



Risks after countermeasures implementation		Severity		
		Low	Medium	High
Probability of occurrence	Low	🟨 Stealing Model Behavior Attack	🟨	🟧 Phishing Attack
	Medium	🟨	🟧	🟥
	High	🟧	🟥	🟥

Fig. 14. Criticality matrix of cyber risks of systems after implementation of countermeasures

Thus, by examining a case study that applied the IMECA framework, we observed how risk-oriented analysis could be effectively utilised within a cyber-physical context. The case study's findings suggest that structured risk assessment enables targeted mitigation strategies, thereby significantly enhancing security. This approach serves as a blueprint for other applications involving AI and advanced technology, where risk assessment is crucial. The insights gained from this case study offer a comprehensive understanding of how to protect complex systems from emerging cyber threats, and they demonstrate the practical value of risk assessment methodologies, such as IMECA.

10. AI challenges & limitations in critical systems

Finalising the analysis of AI security, it is crucial to focus on a wide range of challenges and constraints associated with deploying modern AI in critical systems. This section comprehensively defines both the technical and non-technical aspects of AI integration and deployment, along with an overview of issues related to its application in various fields, such as medicine and military affairs.

Highlighting real challenges and constraints lead to a deeper understanding of potential risks and limitations, thereby ensuring the effective and secure development of this technology in the future. Based on the AI quality model [69], challenges and risks have been classified according to fundamental characteristics. Special attention is given to explainability, lawfulness, and ethics.

10.1. Explainability

The finance and healthcare sectors have increasingly adopted AI technologies to tackle complex tasks and decision-making processes in recent years. For these critical systems, understanding the outcomes and ensuring the confidentiality of information are paramount. Addressing the "black box problem" through

Explainable Artificial Intelligence (XAI) is a significant step towards achieving this goal, as people tend to trust AI more when they can understand its reasoning.

One of the critical challenges in the context of AI is assessing its explainability. Evaluating AI explainability involves developing methods and metrics that gauge how effectively and understandably the system can explain its actions and decisions across different levels [70], balancing explainability with accuracy or performance, particularly in healthcare, where precision is crucial. It's essential to uphold the confidentiality rights of other users. Authors of [71] provide a taxonomy of trends related to explainability techniques for different ML models. Additionally, commercial interests play a significant role in the market. The detailed workings of some of the most widely used machine learning systems, such as Google's search algorithms or language models like ChatGPT, are not publicly accessible to protect competitiveness and intellectual property.

Therefore, the main challenges include defining metrics and methods for assessing AI model explainability to cater to diverse audiences, maintaining a balance between explainability and accuracy, along with the associated costs of ensuring explainability, and ensuring the truthfulness of explanations.

11.2. Lawfulness

Regulating AI presents numerous challenges and risks due to its rapid development and the potential societal implications it poses. One of the primary challenges is safeguarding public interests amidst the AI race. The speed of AI advancements often outpaces existing governmental expertise and regulatory structures, which may not be sufficiently flexible to keep pace with the rapid development of AI. Additionally, regulatory efforts must strike a balanced approach because AI is a multifunctional tool where universal regulations may be excessive or insufficient depending on the use context.

Consideration must also be given to the environmental and climate impacts of AI usage [72]. Machine learning requires vast amounts of data, and the processing and storage of this data consume significant amounts of energy, which in turn impacts the environment and climate. While companies like Google, Amazon, and Microsoft invest in renewable energy and utilise AI to enhance energy efficiency, it remains uncertain whether these investments will sufficiently mitigate the overall global environmental and climate impact of these technologies [72].

Furthermore, the use of AI in military affairs poses significant challenges and risks regarding international law and security. Developing autonomous weapon systems based on AI could potentially violate principles of International Humanitarian Law (LOAC or IHL) [73]. Therefore, key challenges include defining effective mechanisms for monitoring and regulating AI, determining the responsible entities for oversight, and establishing the appropriate levels of control and accountability.

11.3. Ethics

Ethics is defined as that part of philosophy that deals with the prerequisites and evaluation of human action and is the systematic reflection on morality. Ethics is crucial in cases that have a direct impact on human lives. AI discrimination, highlighted by many researchers and governments, refers to preventing bias and injustice caused by AI systems [74]. When AI algorithms process information crucial for human decisions, such as hiring, loan applications, or mortgages, biased data can lead to societal discrimination. Additionally, moral dilemmas arise when AI systems must choose between actions conflicting with moral or ethical values. Authors [75] provide an AI and ethics positioning matrix, indicating that for some AI tools, ethical considerations may be low (e.g., in malware detection). In contrast, for others, it's critical (e.g., intelligent decision support systems in eHealth). They also adapt the PESTEL analysis, where each dimension holds a premise influencing AI's theoretical and hypothetical potential.

Work [76] examines the classification of algorithmic biases, identifying different variants: (1) training data bias could emerge if AI systems are designed with poor, limited, or biased data sets; (2) transfer context bias could emerge when AI systems are designed for one ecological, climate, or social-ecological context and then incorrectly transferred to another; (3) even if the training data and the context in which the algorithm is used are appropriate, their application can still lead to interpretation bias. In this type of bias, an AI system might work as its designer intended. Still, the user does not fully understand its utility or tries to infer different meanings the system might not support.

Therefore, the main challenges include algorithmic biases, discrimination, prejudices, and moral dilemmas [77]. Big data can lead to violations of individual rights and differential treatment, indirectly discriminating against groups with similar characteristics [78, 79]. Thus, the ethics issue involves navigating the trade-off between efficiency and bias.

11. Conclusions

This paper introduces a triadic principle to systematically analyse the roles of AI in cybersecurity, examining it as both an attack tool and a defence mechanism, as well as a system under threat. By mapping out eight distinct scenarios involving combinations of traditional and AI-powered systems, attacks, and protections, the study provides a structured taxonomy for evaluating cyber threats and defence strategies in AI-augmented environments. Through detailed categorisation and analysis of attack types, the study demonstrated that AI not only enhances the sophistication and scale of attacks but also significantly improves detection, prevention, and mitigation capabilities. For each attack, AI-based countermeasures were contrasted with traditional methods, highlighting the superior adaptability and efficiency of ML and DL learning approaches when appropriately designed.

The paper also emphasised the unique vulnerabilities of AI systems themselves, including adversarial manipulation, data poisoning, and model extraction, showing that these systems require dedicated protection mechanisms. The integration of the IMECA methodology for scenario-based risk assessment added a quantitative layer to the evaluation of AI-powered threats, aiding in the prioritisation of risks and the development of mitigation strategies. Despite its comprehensive scope, the study acknowledges that real-world validation and continuous evolution of both attacks and defences remain open challenges. The dynamic nature of AI models and the lack of explainability in some AI-based defences also raise issues for trust and regulatory compliance.

Future research will focus on extending this triadic principle to include adaptive and proactive cyber defence mechanisms, integrating real-time threat intelligence, and addressing the explainability and resilience of AI-driven countermeasures in dynamic environments. Besides, another challenging direction is developing this principle for analysing LLM-systems cybersecurity [80] and sociotechnical attacks [81].

References

1. The Bletchley Declaration by Countries Attending the AI Safety Summit. 2023 (Accessed on 18.07.2024). **https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023/**
2. Veprytska, O., V. Kharchenko. Extended IMECA Technique for Assessing Risks of Successful Cyberattacks. – In: Proc. of 13th International Conference on Dependable Systems, Services and Technologies (DESSERT'23), 2023, pp. 1-7. DOI: 10.1109/DESSERT61349.2023.10416447.1626.
3. Hoque, N., D. K. Bhattacharyya, J. K. Kalita. Botnet in DDoS Attacks: Trends and Challenges. – IEEE Communications Surveys & Tutorials, Vol. **17**, 2015, pp. 2242-2270. DOI: 10.1109/comst.2015.2457491.
4. Xing, Y., H. Shu, H. Zhao, D. Li, L. Guo. Survey on Botnet Detection Techniques: Classification, Methods, and Evaluation. – Mathematical Problems in Engineering, Vol. **2021**, 2021, 6640499. DOI: 10.1155/2021/6640499.
5. Somani, G., M. S. Gaur, D. Sanghi, M. Conti, R. Buyya. DDoS Attacks in Cloud Computing: Issues, Taxonomy, and Future Directions. – Computer Communications, Vol. **107**, 2017, pp. 30-48. DOI: 10.1016/j.comcom.2017.03.010.
6. Vishwakarma, R., A. K. Jain. A Survey of DDoS Attacking Techniques and Defence Mechanisms in the IoT Network. – Telecommunication Systems, Vol. **73**, 2020, pp. 3-25. DOI: 10.1007/s11235-019-00599-z.
7. Bhardwaj, A., V. Mangat, R. Vig, S. Halder, M. Conti. Distributed Denial of Service Attacks in Cloud: S tate-of-the-Art of Scientific and Commercial Solutions. – Computer Science Review, Vol. **39**, 2021, 100332. DOI: 10.1016/j.cosrev.2020.100332.
8. Jaafar, G. A., S. M. Abdullah, S. Ismail. Review of Recent Detection Methods for HTTP DDoS Attack. – Journal of Computer Networks and Communications, Vol. **2019**, 2019, 1283472. DOI: 10.1155/2019/1283472.
9. Singh, A., B. B. Gupta. Distributed Denial-of-Service (DDoS) Attacks and Defence Mechanisms in Various Web-Enabled Computing Platforms: Issues, Challenges, and Future Research Directions. – International Journal on Semantic Web and Information Systems (IJSWIS), Vol. **18**, 2022, pp. 1-43. DOI: 10.4018/ijswis.297143.
10. Dhadhal, H., P. Kotak. Leveraging Datasets for Effective Mitigation of DDoS Attacks in Software-Defined Networking: Significance and Challenges. – Radioelectronic and Computer Systems, 2024, No 2, pp. 136-146. DOI: 10.32620/reks.2024.2.11.
11. Mittal, M., K. Kumar, S. Behal. Deep Learning Approaches for Detecting DDoS Attacks: A Systematic Review. – Soft Computing, Vol. **27**, 2023, pp. 13039-13075. DOI: 10.1007/s00500-021-06608-1.
12. Sood, A. K., S. Zeadally. A Taxonomy of Domain-Generation Algorithms. – IEEE Security & Privacy, Vol. **14**, 2016, pp. 46-53. DOI: 10.1109/msp.2016.76.

13. Charan, P. S., S. K. Shukla, P. M. Anand. Detecting Word Based DGA Domains Using Ensemble Models. – In: Proc. of 19th International Conference of Cryptology and Network Security (CANS'20), Vienna, Austria, 14-16 December 2020, Proceedings, 19. Springer, 2020, pp. 127-143. DOI: 10.1007/978-3-030-65411-5_7.

14. Mao, J., J. Zhang, Z. Tang, Z. Gu. DNS Anti-Attack Machine Learning Model for DGA Domain Name Detection. – Physical Communication, Vol. **40**, 2020, 101069. DOI: 10.1016/j.phycom.2020.101069.

15. Chiscop, I., F. Soro, P. Smith. AI-Based Detection of DNS Misuse for Network Security. – In: Proc. of 1st International Workshop on Native Network Intelligence, 2022, pp. 27-32. DOI: 10.1145/3565009.3569523.

16. Shu, K., A. Sliva, S. Wang, J. Tang, H. Liu. Fake News Detection on Social Media: A Data Mining Perspective. – ACM SIGKDD Explorations Newsletter, Vol. **19**, 2017, pp. 22-36. DOI: 10.1145/3137597.3137600.

17. Padalko, H., V. Chomko, S. Yakovlev, D. Chumachenko. Ensemble Machine Learning Approaches for Fake News Classification. – Radioelectronic and Computer Systems, 2023, No 4, pp. 5-19. DOI: 10.32620/reks.2023.4.01.

18. Zhang, X., A. A. Ghorbani. An Overview of Online Fake News: Characterization, Detection, and Discussion. – Information Processing & Management, Vol. **57**, 2020, 102025. DOI: 10.1016/j.ipm.2019.03.004.

19. Zhou, X., R. Zafarani. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. – ACM Computing Surveys (CSUR), Vol. **53**, 2020, pp. 1-40. DOI: 10.1145/3395046.

20. Tandoc Jr, E. C., Z. W. Lim, R. Ling. Defining "Fake News" A Typology of Scholarly Definitions. – Digital Journalism, Vol. **6**, 2018, pp. 137-153. DOI: 10.1080/21670811.2017.1360143.

21. Bharati, A., D. Moreira, J. Brogan, P. Hale, K. Bowyer, P. Flynn, A. Rocha, W. Scheirer. Beyond Pixels: Image Provenance Analysis Leveraging Metadata. – In: Proc. of IEEE Winter Conference on Applications of Computer Vision (WACV'19), IEEE, 2019, pp. 1692-1702. DOI: 10.1109/wacv.2019.00185.

22. Vadrevu, A., R. Rajeshwari, L. Pabbathi, S. Sirimalla, D. Vodnala. Image Forgery Detection Using Metadata Analysis and ELA Processor. – In: Innovations in Computer Science and Engineering: Proceedings of the Ninth ICICSE, 2021, Springer, 2022, pp. 579-586. DOI: 10.1007/978-981-16-8987-1_62.

23. Fischinger, D., M. Boyer. DF2023: The Digital Forensics 2023 Dataset for Image Forgery Detection. In Proceedings of the DF2023: The Digital Forensics 2023 Dataset for Image Forgery Detection. – In: P. Corcoran, Ed. Proc. of 25th Irish Machine Vision and Image Processing Conference (IMVIP'23), 2023, Through 01-09-2023, pp. 128-135.

24. Liu, Z., X. Qi, P. H. Torr. Global Texture Enhancement for Fake Face Detection. – In: Proc. of Conf. on Computer Vision & Pattern Recognition (IEEE/CVF'20), 2020.

25. Wang, C., W. Deng. Representative Forgery Mining for Fake Face Detection. – In: Proc. of Conf. on Computer Vision & Pattern Recognition (IEEE/CVF'21), 2021.

26. Westerlund, M. The Emergence of Deepfake Technology: A Review. – Technology Innovation Management Review, Vol. **9**, 2019.

27. Yu, P., Z. Xia, J. Fei, Y. Lu. A Survey on Deepfake Video Detection. – Iet Biometrics, Vol. **10**, 2021, pp. 607-624. DOI: 10.1049/bme2.12031.

28. Badale, A., L. Castelino, C. Darekar, J. Gomes. Deep Fake Detection Using Neural Networks. – In: Proc. of 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS'18), Vol. **2**, 2018.

29. Güera, D., E. J. Delp. Deepfake Video Detection Using Recurrent Neural Networks. – In: Proc. of 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS'18), IEEE, 2018, pp. 1-6. DOI: 10.1109/avss.2018.8639163.

30. Alkhalil, Z., C. Hewage, L. Nawaf, I. Khan. Phishing Attacks: A Recent Comprehensive Study and a New Anatomy. – Frontiers in Computer Science, Vol. **3**, 2021, 563060. DOI: 10.3389/fcomp.2021.563060.

31. Rastenis, J., S. Ramanauskaitė, J. Janulevičius, A. Čenys, A. Slotkienė, K. Pakrijauskas. e-Mail-Based Phishing Attack Taxonomy. – Applied Sciences, Vol. **10**, 2020, 2363. DOI: 10.3390/app10072363.

32. Alabdan, R. Phishing Attacks Survey: Types, Vectors, and Technical Approaches. – Future Internet, Vol. **12**, 2020, 168. DOI: 10.3390/fi12100168.

33. Güera, D., E. J. Delp. Deepfake Video Detection Using Recurrent Neural Networks. – In: Proc. of 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS'18), IEEE, 2018, pp. 1-6. DOI: 10.1109/avss.2018.8639163.

34. Vijayalakshmi, M., S. Mercy Shalinie, M. H. Yang, R. M. U. Web Phishing Detection Techniques: A Survey on the State-of-the-Art, Taxonomy and Future Directions. – Iet Networks, Vol. **9**, 2020, pp. 235-246. DOI: 10.1049/iet-net.2020.0078.

35. Basit, A., M. Zafar, X. Liu, A. R. Javed, Z. Jalil, K. Kifayat. A Comprehensive Survey of AI-Enabled Phishing Attacks Detection Techniques. – Telecommunication Systems,
Vol. **76**, 2021, pp. 139-154. DOI: 10.1007/s11235-020-00733-2.

36. Shahrivari, V., M. M. Darabi, M. Izadi. Phishing Detection Using Machine Learning Techniques. – arXiv Preprint arXiv:2009.11116 2020.

37. Do, N. Q., A. Selamat, O. Krejcar, E. Herrera-Viedma, H. Fujita. Deep Learning for Phishing Detection: Taxonomy, Current Challenges and Future Directions. – IEEE Access, Vol. **10**, 2022, pp. 36429-36463.

38. Shirazi, H., S. R. Muramudalige, I. Ray, A. P. Jayasumana. Improved Phishing Detection Algorithms Using Adversarial Autoencoder Synthesized Data. – In: Proc. of 45th IEEE Conference on Local Computer Networks (LCN'20). IEEE, 2020, pp. 24-32. DOI: 10.1109/lcn48667.2020.9314775.

39. Yu, G., W. Fan, W. Huang, J. An. An Explainable Method of Phishing Email Generation and Its Application in Machine Learning. – In: Proc. of 4th IEEE Information Technology, Networking, Electronic and Automation Control Conference (ITNEC'20). IEEE, 2020, Vol. **1**, pp. 1279-1283. DOI: 10.1109/itnec48623.2020.9085171.

40. Sahingoz, O. K., E. Buber, O. Demir, B. Diri. Machine Learning Based Phishing Detection from URLs. – Expert Systems with Applications, Vol. **117**, 2019, pp. 345-357. DOI: 10.1016/j.eswa.2018.09.029.

41. Guembe, B., A. Azeta, S. Misra, V. C. Osamor, L. Fernandez-Sanz, V. Pospelova. The Emerging Threat of AI-Driven Cyber Attacks: A Review. – Applied Artificial Intelligence, Vol. **36**, 2022, 2037254. DOI: 10.1080/08839514.2022.2037254.

42. Cinà, A. E., K. Grosse, A. Demontis, S. Vascon, W. Zellinger, B. A. Moser, A. Oprea, B. Biggio, M. Pelillo, F. Roli. Wild Patterns Reloaded: A Survey of Machine Learning Security against Training Data Poisoning. – ACM Computing Surveys, Vol. **55**, 2023, pp. 1-39. DOI: 10.1145/3585385.

43. Ramirez, M. A., S. K. Kim, H. A. Hamadi, E. Damiani, Y. J. Byon, T. Y. Kim, C. S. Cho, C. Y. Yeun. Poisoning Attacks and Defences on Artificial Intelligence: A Survey. – arXiv Preprint arXiv:2202.10276 2022. 1749.

44. Gao, Y., B. G. Doan, Z. Zhang, S. Ma, J. Zhang, A. Fu, S. Nepal, H. Kim. Backdoor Attacks and Countermeasures on Deep Learning: A Comprehensive Review. – arXiv Preprint arXiv:2007.107602020.

45. Li, Y., Y. Jiang, Z. Li, S. T. Xia. Backdoor Learning: A Survey. – IEEE Transactions on Neural Networks and Learning Systems, Vol. **35**, 2022, pp. 5-22.

46. Xia, G., J. Chen, C. Yu, J. Ma. Poisoning Attacks in Federated Learning: A Survey. – IEEE Access, Vol. **11**, 2023, pp. 10708-10722. DOI: 10.1109/ACCESS.2023.3238823.

47. Sikandar, H. S., H. Waheed, S. Tahir, S. U. Malik, W. Rafique. A Detailed Survey on Federated Learning Attacks and Defences. – Electronics, Vol. **12**, 2023, 260. DOI: 10.3390/electronics12020260.

48. Rodríguez-Barroso, N., D. Jiménez-López, M. V. Luzón, F. Herrera, E. Martínez-Cámara. Survey on Federated Learning Threats: Concepts, Taxonomy on Attacks and Defences, Experimental Study and Challenges. – Information Fusion, Vol. **90**, 2023, pp. 148-173.

49. Lu, Y. G., Kamath, Y. Yu. Indiscriminate Data Poisoning Attacks on Neural Networks. – arXiv Preprint arXiv:2204.09092 2022.

50. Suciu, O., R. Marginean, Y. Kaya, H. Daume III, T. Dumitras. When Does Machine Learning {FAIL}? Generalized Transferability for Evasion and Poisoning Attacks. – In: Proc. of 27th USENIX Security Symposium, 2018, pp. 1299-1316.

51. Vassilev, A., A. Oprea, A. Fordyce, H. Anderson. Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. Technical Report, National Institute of Standards and Technology, 2024. DOI: 10.6028/nist.ai.100-2e2023.

52. Oliynyk, D., R. Mayer, A. Rauber. I Know What You Trained Last Summer: A Survey on Stealing Machine Learning Models and Defences. – ACM Computing Surveys, Vol. **55**, 2023, pp. 1-41. DOI: 10.1145/3595292.

53. Rigaki, M., S. Garcia. A Survey of Privacy Attacks in Machine Learning. – ACM Computing Surveys, Vol. **56**, 2023, pp. 1-34. DOI: 10.1145/3624010.

54. Liu, X., L. Xie, Y. Wang, J. Zou, J. Xiong, Z. Ying, A. V. Vasilakos. Privacy and Security Issues in Deep Learning: A Survey. – IEEE Access, Vol. **9**, 2021, pp. 4566-4593. DOI: 10.1109/ACCESS.2020.3045078.

55. Liu, B., M. Ding, S. Shaham, W. Rahayu, F. Farokhi, Z. Lin. When Machine Learning Meets Privacy: A Survey and Outlook. – ACM Computing Surveys (CSUR), Vol. **54**, 2021, pp. 1-36. DOI: 10.1145/3436755.

56. Miao, Y., C. Chen, L. Pan, Q. L. Han, J. Zhang, Y. Xiang. Machine Learning-Based Cyber Attacks Targeting on Controlled Information: A Survey. – ACM Computing Surveys (CSUR), Vol. **54**, 2021, pp. 1-36.

57. Xiang, Y., Z. Chen, Z. Chen, Z. Fang, H. Hao, J. Chen, Y. Liu, Z. Wu, Q. Xuan, X. Yang. Open DNN Box by Power Side-Channel Attack. – IEEE Transactions on Circuits and Systems II: Express Briefs, Vol. **67**, 2020, pp. 2717-2721.

58. Carlini, N. A Complete List of All (arXiv) Adversarial Example Papers (Accessed on 18.07.2024).
https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html

59. Ren, K., T. Zheng, Z. Qin, X. Liu. Adversarial Attacks and Defences. – Deep Learning. Engineering, Vol. **6**, 2020, pp. 346-360. DOI: 10.1016/j.eng.2019.12.012.

60. Deng, Y., X. Zheng, T. Zhang, C. Chen, G. Lou, M. Kim. An Analysis of Adversarial Attacks and Defences on Autonomous Driving Models. – In: Proc. of IEEE International Conference on Pervasive Computing and Communications (PerCom'20), IEEE, 2020, pp. 1-10. DOI: 10.1109/PerCom45495.2020.9127389.

61. Liang, H., E. He, Y. Zhao, Z. Jia, H. Li. Adversarial Attack and Defence: A Survey. – Electronics, Vol. **11**, 2022, 1283. DOI: 10.3390/electronics11081283.

62. Wang, Y., T. Sun, S. Li, X. Yuan, W. Ni, E. Hossain, H. V. Poor. Adversarial Attacks and Defences in Machine Learning-Empowered Communication Systems and Networks: A Contemporary Survey. – IEEE Communications Surveys & Tutorials, 2023.

63. Brown, T. B., D. Mané, A. Roy, M. Abadi, J. Gilmer. Adversarial Patch. – arXiv Preprint arXiv:1712.09665 2017.

64. Ren, K., T. Zheng, Z. Qin, X. Liu. Adversarial Attacks and Defences. – Deep Learning. Engineering, Vol. **6**, 2020, pp. 346-360. DOI: 10.1016/j.eng.2019.12.012.

65. Shumailov, I., Y. Zhao, D. Bates, N. Papernot, R. Mullins, R. Anderson. Sponge Examples: Energy-Latency Attacks on Neural Networks. – In: Proc. of IEEE European Symposium on Security and Privacy (EuroS&P'21), IEEE, 2021, pp. 212-231.

66. Zouave, E., M. Bruce, K. Colde, M. Jaitner, I. Rodhe, T. Gustafsson. Artificially Intelligent Cyberattacks. Stockholm: Totalförsvarets Forskningsinstitut FOI (Accessed on 18.08.2024).
https://www.statsvet.uu.se/digitalAssets/769/c_769530-l_3-k_rapport-foi-vt20.pdf

67. EU Action Plan against Disinformation, 2020 (Accessed on 18.08.2024).
https://www.eeas.europa.eu/sites/default/files/action_plan_against_disinformation.pdf

68. 2024 State of the Phish – Today's Cyber Threats and Phishing Protection, 2024 (Accessed on 18.08.2024).
https://www.proofpoint.com/us/resources/threat-reports/state-of-phish

69. Kharchenko, V., H. Fesenko, O. Illiashenko. Quality Models for Artificial Intelligence Systems: Characteristic-Based Approach. – Development and Application. Sensors, Vol. **22**, 2022, 4865. DOI: 10.3390/s22134865.

70. Younis, H. A., T. A. E. Eisa, M. Nasser, T. M. Sahib, A. A. Noor, O. M. Alyasiri, S. Salisu, I. M. Hayder, H. A. Younis. A Systematic Review and Meta-Analysis of Artificial Intelligence Tools in Medicine and Healthcare: Applications, Considerations, Limitations, Motivation and Challenges. – Diagnostics, Vol. **14**, 2024, 109. DOI: 10.3390/diagnostics14010109.

71. Arrieta, A. B., N. Díaz-Rodríguez, J. del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. – Information Fusion, Vol. **58**, 2020, pp. 82-115. DOI: 10.1016/j.inffus.2019.12.012.

72. Coeckelbergh, M. AI for Climate: Freedom, Justice, and Other Ethical and Political Challenges. – AI and Ethics, Vol. **1**, 2021, pp. 67-72. DOI: 10.1007/s43681-020-00007-2.

73. Forrest, E. M., B. Boudreaux. Military Applications of Artificial Intelligence, 2020.

74. Wirtz, B. W., J. C. Weyerer, B. J. Sturm. The Dark Sides of Artificial Intelligence: An Integrated AI Governance Framework for Public Administration. – International Journal of Public Administration, Vol. **43**, 2020, pp. 818-829. DOI: 10.1080/01900692.2020.1749851.

75. Brendel, A. B., M. Mirbabaie, T. B. Lembcke, L. Hofeditz. Ethical Management of Artificial Intelligence. – Sustainability, Vol. **13**, 2021, 1974. DOI: 10.3390/su13041974.

76. Galaz, V., M. A. Centeno, P. W. Callahan, A. Causevic, T. Patterson, I. Brass, S. Baum, D. Farber, J. Fischer, D. Garcia et al. Artificial Intelligence, Systemic Risks, and Sustainability. – Technology in Society, Vol. **67**, 2021, 101741. DOI: 10.1016/j.techsoc.2021.101741.

77. Edris, E. K. K. Utilisation of Artificial Intelligence and Cybersecurity Capabilities: A Symbiotic Relationship for Enhanced Security and Applicability. – Electronics, Vol. **14**, 2025, 2057. DOI: 10.3390/electronics14102057.

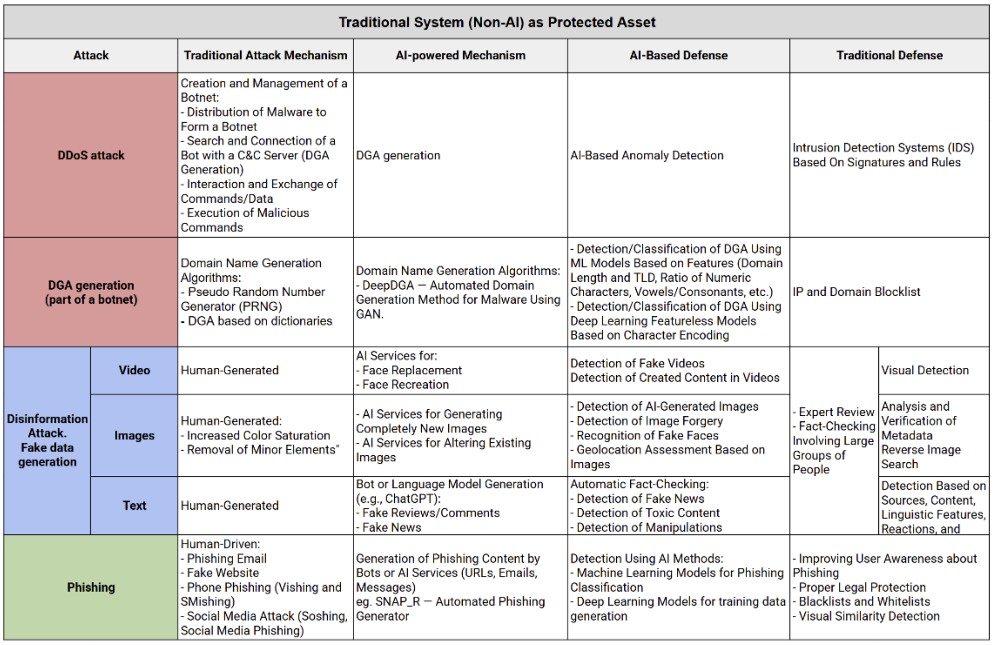
78. Rodrigues, R. Legal and Human Rights Issues of AI: Gaps, Challenges and Vulnerabilities. – Journal of Responsible Technology, Vol. **4**, 2020, 100005. DOI: 10.1016/j.jrt.2020.100005.

79. Okdem, S., S. Okdem. Artificial Intelligence in Cybersecurity: A Review and a Case Study. – Appl. Sci., Vol. **14**, 2024, 10487. DOI: 10.3390/app142210487.

80. Neretin, O., V. Kharchenko. A Model of Ensuring LLM Cybersecurity. – Radioelectronic and Computer Systems, 2025, No 2, pp. 201-215. DOI: 10.32620/reks.2025.2.13.

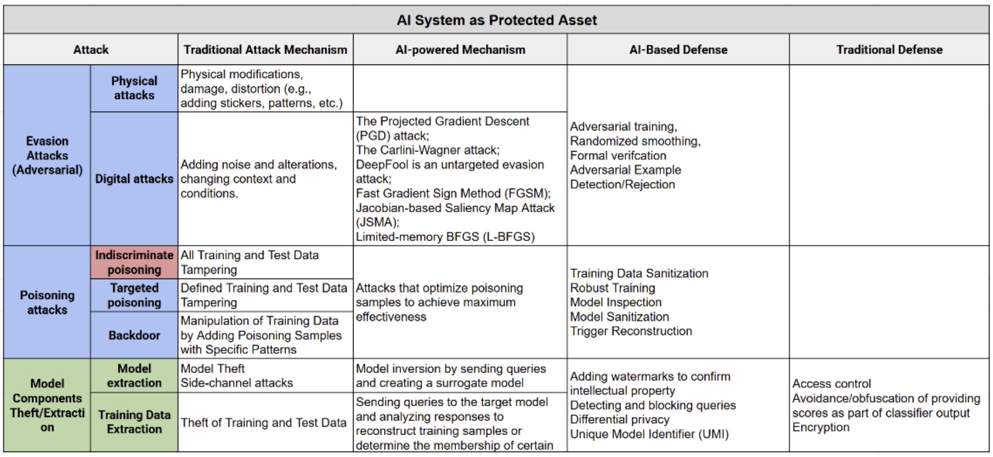
81. Korchenko, O., A. Korchenko, S. Zybin, K. Davydenko. An Approach for Classifying Sociotechnical Attacks. – Radioelectronic and Computer Systems, 2025, No 2, pp. 230-252. DOI: 10.32620/reks.2025.2.15.

Appendix A. Classification of Countermeasures against Various Attack Scenarios Based on Their Impact on Availability (Red), Integrity (Blue), and Confidentiality (Green) on Traditional Systems



Traditional System (Non-AI) as Protected Asset						
Attack		Traditional Attack Mechanism	AI-powered Mechanism	AI-Based Defense	Traditional Defense	
DDoS attack		Creation and Management of a Botnet: - Distribution of Malware to Form a Botnet - Search and Connection of a Bot with a C&C Server (DGA Generation) - Interaction and Exchange of Commands/Data - Execution of Malicious Commands	DGA generation	AI-Based Anomaly Detection	Intrusion Detection Systems (IDS) Based On Signatures and Rules	
DGA generation (part of a botnet)		Domain Name Generation Algorithms: - Pseudo Random Number Generator (PRNG) - DGA based on dictionaries	Domain Name Generation Algorithms: - DeepDGA — Automated Domain Generation Method for Malware Using GAN.	- Detection/Classification of DGA Using ML Models Based on Features (Domain Length and TLD, Ratio of Numeric Characters, Vowels/Consonants, etc.) - Detection/Classification of DGA Using Deep Learning Featureless Models Based on Character Encoding	IP and Domain Blocklist	
Disinformation Attack. Fake data generation	Video	Human-Generated	AI Services for: - Face Replacement - Face Recreation	Detection of Fake Videos Detection of Created Content in Videos	- Expert Review - Fact-Checking Involving Large Groups of People	Visual Detection
	Images	Human-Generated: - Increased Color Saturation - Removal of Minor Elements"	- AI Services for Generating Completely New Images - AI Services for Altering Existing Images	- Detection of AI-Generated Images - Detection of Image Forgery - Recognition of Fake Faces - Geolocation Assessment Based on Images		Analysis and Verification of Metadata Reverse Image Search
	Text	Human-Generated	Bot or Language Model Generation (e.g., ChatGPT): - Fake Reviews/Comments - Fake News	Automatic Fact-Checking: - Detection of Fake News - Detection of Toxic Content - Detection of Manipulations		Detection Based on Sources, Content, Linguistic Features, Reactions, and
Phishing		Human-Driven: - Phishing Email - Fake Website - Phone Phishing (Vishing and SMishing) - Social Media Attack (Soshing, Social Media Phishing)	Generation of Phishing Content by Bots or AI Services (URLs, Emails, Messages) eg. SNAP_R — Automated Phishing Generator	Detection Using AI Methods: - Machine Learning Models for Phishing Classification - Deep Learning Models for training data generation	- Improving User Awareness about Phishing - Proper Legal Protection - Blacklists and Whitelists - Visual Similarity Detection	

Appendix B. Classification of Countermeasures against Various Attack Scenarios Based on Their Impact on Availability (Red), Integrity (Blue), and Confidentiality (Green) on AI Systems



AI System as Protected Asset					
Attack		Traditional Attack Mechanism	AI-powered Mechanism	AI-Based Defense	Traditional Defense
Evasion Attacks (Adversarial)	Physical attacks	Physical modifications, damage, distortion (e.g., adding stickers, patterns, etc.)		Adversarial training, Randomized smoothing, Formal verifcation Adversarial Example Detection/Rejection	

	Digital attacks	Adding noise and alterations, changing context and conditions.	The Projected Gradient Descent (PGD) attack;
			The Carlini-Wagner attack;
			DeepFool is an untargeted evasion attack;
			Fast Gradient Sign Method (FGSM);
			Jacobian-based Saliency Map Attack (JSMA);
			Limited-memory BFGS (L-BFGS)		

Poisoning attacks	Indiscriminate poisoning	All Training and Test Data Tampering	Attacks that optimize poisoning samples to achieve maximum effectiveness	Training Data Sanitization Robust Training Model Inspection Model Sanitization Trigger Reconstruction	
	Targeted poisoning	Defined Training and Test Data Tampering			
	Backdoor	Manipulation of Training Data by Adding Poisoning Samples with Specific Patterns			
Model Components Theft/Extraction	Model extraction	Model Theft Side-channel attacks	Model inversion by sending queries and creating a surrogate model	Adding watermarks to confirm intellectual property Detecting and blocking queries Differential privacy Unique Model Identifier (UMI)	Access control Avoidance/obfuscation of providing scores as part of classifier output Encryption
	Training Data Extraction	Theft of Training and Test Data	Sending queries to the target model and analyzing responses to reconstruct training samples or determine the membership of certain		

Fast-track. Received: 07.08.2025. Revised Version: 27.08.2025. Accepted: 01.09.2025