TECHNOLOGIES AND ENGINEERING

Journal homepage: https://technologies-engineering.com.ua/en
Vol. 26, No. 4, 2025

Received: 22.04.2025
Revised: 18.08.2025
Accepted: 17.09.2025

UDC 004.056.5:004.8

DOI: 10.30857/2786-5371.2025.4.11

AI vs AI: Using artificial intelligence to protect data and attacks on it

Igor Rudenko*
Master
Yaroslav Mudryi National Law University
61024, 77 Hryhorii Skovoroda Str., Kharkiv, Ukraine
https://orcid.org/0009-0008-3582-3951

Abstract. The aim of the study was to conduct a comprehensive analysis of the role of artificial intelligence as a tool for attacks and defence in information systems, with a focus on evaluating the effectiveness of existing approaches and justifying prospects for integrating artificial intelligence to enhance cybersecurity under conditions of increasing intelligent threats. Within the framework of the study, the confrontation between offensive and defensive artificial intelligence systems in a dynamic environment with adaptive behaviour was modelled, which made it possible not only to identify typical threat vectors but also to assess the effectiveness of corresponding countermeasures. It was found that generative models, particularly those based on reinforcement learning, effectively adapted to defensive responses, bypassing traditional filters and heuristics. At the same time, the highest resilience to such attacks was demonstrated by combined approaches that integrated Federated Learning, blockchain, and differential privacy: the level of resistance to attacks increased by up to 40% with a moderate decrease in accuracy (3-6%). Adversarial training ensured an increase in security of up to 25%, although the accuracy dropped by up to 4%, and its effectiveness depended significantly on the completeness and variability of training data. Homomorphic encryption proved to be the most confidential approach, but remained limited in practical use due to excessive resource consumption and processing time. Although blockchain tools contributed to transparency and data immutability, these tools involved high latency, which complicated the application in real-time conditions. Overall, the results of the study confirmed the appropriateness of using multimodal, adaptive, and multi-level protection strategies for artificial intelligence systems, especially amid the growing number of generative attacks, as evidenced by real cases (e.g., Disney). The practical significance lay in the formation of foundations for the development of adaptive cyber defence systems capable of countering intelligent attacks in real time. The obtained results could be used to enhance the security of critical infrastructure, financial platforms, and autonomous systems

Keywords: data privacy; generative networks; phishing; biometric system; deepfakes

Introduction

The relevance of the study stemmed from the rapid development of artificial intelligence (AI), which played a key role in modern cybersecurity systems. However, alongside this, the use of AI for attacks also increased, significantly complicating traditional methods of information protection. Modern offensive systems, due to the ability to adapt and evolve in real time, could bypass traditional protection methods, change malware signatures, create phishing messages, or attack biometric systems. This created new challenges for cybersecurity professionals and required the development of innovative approaches to designing defensive systems capable of independently adapting to new threats. The "AI vs AI" system had become a reality, where attacks and defence continuously interacted, learning from each other – presenting researchers with the task of creating AI capable of effectively countering hostile intelligent systems. This study considered the effectiveness of such interactions and proposed new strategies for developing defensive systems that could form the foundation for enhancing cybersecurity amid escalating technological threats.

During the period 2020-2025, the topic of "AI vs AI" confrontation attracted the attention of many researchers,

Suggested Citation:
Rudenko, I. (2025). AI vs AI: Using artificial intelligence to protect data and attacks on it. *Technologies and Engineering,* 26(4), 11-25. doi: 10.30857/2786-5371.2025.4.11.

*Corresponding author

who proposed various approaches to building both defensive and offensive systems based on AI. To form a comprehensive view of this issue, the conditional findings of ten authors whose research was considered relevant to the selected topic were analysed. G. Pu *et al.* (2020) focused on developing a hybrid intrusion detection system based on a combination of unsupervised learning and clustering methods. The authors showed that such a system detected new types of attacks but faced the problem of a high number of false positives when the network topology changed. S. Al-Ahmadi *et al.* (2022) investigated the use of Generative Adversarial Network (GAN) for generating phishing emails capable of bypassing modern anti-phishing filters. The results demonstrated the ability of AI-attacks to outperform traditional filtering tools, highlighting the need for new forms of dynamic content analysis. D. Calvaresi *et al.* (2021) proposed the use of reinforcement learning agents for continuous real-time system monitoring and demonstrated the effectiveness of this approach under changing input data, but noted that the models required significant training and computational resources. R.R. Gopireddy (2024) analysed data poisoning attacks, where the attacker manipulated training sets to distort the performance of the defensive AI. The results showed that even slight data distortion significantly reduced the accuracy of deep learning-based models.

J. Wu *et al.* (2022) presented a multi-level defence system based on an ensemble of models that intercepted complex multistep attacks. At the same time, the authors pointed out the risks of excessive complexity, which complicated the interpretability of model decisions. D.L. Marino *et al.* (2025) proposed the use of transformer architectures to analyse network traffic for anomaly detection. The approach demonstrated high accuracy but remained vulnerable to attacks that mimicked legitimate user behaviour. A. Ragmani *et al.* (2020) studied the effectiveness of adaptive neural network-based models for protecting cloud services. The research showed that such models coped better than traditional ones with Distributed Denial of Service (DDoS) attacks, but worked less effectively under limited input data. K. Santhi *et al.* (2024) conducted an analysis of ethical risks associated with the use of AI in defensive systems, particularly the likelihood of false user identification or breaches of confidentiality. The authors emphasised the need to ensure algorithmic transparency during implementation. M. Hasan (2024) explored the integration of defensive models with blockchain technologies to counter event log modification in cyberinfrastructure. The solution demonstrated resistance to manipulation but required optimisation for transaction processing delays. P.S. Muhuri *et al.* (2020) developed an attack vector prediction system based on historical data and deep recurrent networks. The approach reduced incident response time by 30%, but proved less effective under conditions of data scarcity or non-representativeness.

Thus, the summarised analysis showed that no approach was universal. The development of the "AI vs AI" direction required not only engineering innovations but also a deep interdisciplinary understanding in the fields of machine learning, ethics, security, and information technology. Despite significant progress in applying AI to protect information systems, a number of important aspects remained insufficiently researched. In particular, most existing approaches focused on the isolated use of AI either for attacks or for defence, whereas the dynamic interaction between offensive and defensive AI systems in real time remained limited in study. Insufficient attention was also paid to the resilience of defensive models in the case of combined intelligent attacks involving multiple tactics – for example, data poisoning alongside phishing or social engineering. Furthermore, the question of defensive systems' adaptability to rapidly evolving threats posed by self-learning attack algorithms remained open. All of this determined the need for a deeper study of the "AI vs AI" confrontation in the context of the dynamic and complex environment of modern cybersecurity.

The aim of the study was to determine the effectiveness and limitations of modern AI-based defence tools in the context of evolutionary confrontation with AI attacks, with the further justification of directions for improving the cyber resilience of information systems. To achieve this aim, the following tasks were set: to model typical interaction scenarios between defensive and offensive AI, identify critical vulnerabilities in defensive model performance under intelligent attacks, and propose approaches to increasing the adaptability and effectiveness of defensive systems in a constantly changing threat environment.

Materials and Methods

This study implemented a multi-level approach to the analysis and modelling of the confrontation between offensive and defensive AI under conditions of escalating cyber threats and increasing complexity of digital infrastructures. The primary objective was to build a conceptual and practical model of interaction between AI systems within the "AI vs AI" scenario, reflecting the realities of modern cyberspace where both sides of the conflict could employ intelligent agents. The study analysed scientific publications, real-world attack cases, and statistical data covering the period from 2020 to March 2025. This timeframe made it possible to trace the evolution of offensive and defensive strategies of AI systems, as well as to monitor changes in the frequency, complexity, and effectiveness of security incidents. The analysis used sources from academic publications, technical documentation, cybersecurity reports, and open cases of digital infrastructure breaches.

Particular attention was paid to generative attacks, especially face falsification using GANs, in the context of the increasing difficulty of detecting forgeries in biometric systems. Physical objects such as glasses with modified textures were considered as potential tools for carrying out subtle attacks on facial recognition systems. In the field of attacks on text-based systems, the use of stylometric analysers and filters focused on detecting AI-generated

phishing messages was analysed. The study also covered novel approaches to model inversion based on facial reconstruction from latent space using diffusion mechanisms. Among the examined countermeasures was the use of federated learning to reduce data compromise risks through decentralised processing. In the cryptographic domain, the application of fully homomorphic encryption was analysed, allowing computations to be performed on encrypted data without decryption. Additionally, in the context of dynamic deployment of security infrastructure, the use of unmanned aerial vehicles as mobile nodes capable of adaptively responding to threats in real time was explored.

The first stage of the study involved developing a classification system for offensive AI agents that enabled a structured representation of the functionality, goals, adaptability, and level of autonomy. The classification considered four key parameters: type of action (analytical, generative, manipulative), training model (static training, continuous learning, reinforcement learning), attack target (data, infrastructure, user), and level of autonomy (human-operated, semi-autonomous, fully autonomous agents). This made it possible to identify the most high-risk combinations of traits – for example, fully autonomous agents with generative functions that adapted to environmental changes in real time. The second stage included practical modelling and simulation of interaction scenarios between AI agents on both sides. Frameworks such as TensorFlow 2.0, PyTorch, and the OpenAI Gym environment were used to implement the models. The agents were implemented as modules with clearly distributed functions: offensive agents generated penetration scenarios, data manipulation, or deception model training, while defensive agents performed detection, classification, blocking, or recovery. Particular attention was paid to adversarial examples, where minimal changes in input data could result in misclassification. The third stage involved a comparative analysis of simulation results. The effectiveness of system interaction was assessed using several criteria: average attack detection time, classification accuracy, false positive rate, model adaptability, and computational costs for both sides. The obtained data made it possible to identify weak points in defence systems under conditions of dynamic offensive strategies. Real-world cases were also used to build simulations. Thus, the research methodology combined theoretical generalisation with practical modelling, providing in-depth understanding of the characteristics of the confrontation between intelligent systems under next-generation cyber threats.

Results

Classification of AI attack tools and defence strategies in a dynamic environment. In the context of the "AI vs AI" confrontation study, it is appropriate to classify AI tools used for attacks according to several criteria: type of action, training method, usage purpose, and level of autonomy. Depending on the type of action, offensive AI was categorised as information-analytical, generative, and aggressively influential. Information-analytical agents were used to collect and process data on the target, identify vulnerabilities, and analyse user behaviour, particularly through the use of natural language processing to scan open sources. Generative models, including GANs, enabled the creation of fake images, videos, messages, and other content used in phishing campaigns or social engineering attacks. Aggressively influential models were aimed at actively disrupting systems – through injecting malicious data into training sets, generating input data to deceive the model, or modifying data during transmission.

By training method, these AI systems were divided into pre-trained and adaptive. The former relied on a pre-formed knowledge base and were designed for mass attacks, while the latter learned in real time, adjusting the behaviour based on the defence system's responses, increasing the attack's effectiveness. According to the usage purpose, attacks targeted data, models, or users. Data attacks involved stealing, encrypting, or destroying information using intelligent analysis of file structures and access rights. Model attacks aimed at reverse engineering or identifying weaknesses in the defence system's logic. In turn, user-targeted attacks involved imitating human behaviour, personalising phishing messages, or manipulative communication based on psychological profiling (Karl & Potter, 2023).

AI-based attacks and defences exist in constant dynamic – offensive AI analyses vulnerabilities in defensive systems, attempts to bypass protective measures, and finds new paths for intrusion, while defensive AI, in turn, adapts to these changes. For instance, when generative models (like GANs) are used to create fake data, a defensive system may detect such data through anomaly pattern analysis, making it increasingly difficult over time for offensive models to succeed. However, defence systems must respond to these changes in time, otherwise an attack may succeed. This battle is often cyclical: attacks evolve, and the discovery of new vulnerabilities or strategies is continuously followed by improved defence models (Kumar *et al.*, 2024). A distinctive feature of the interaction between offensive and defensive AI is the adaptability of both sides. In the modern environment, where offensive systems can actively adjust to new conditions and defence strategies, defensive AI systems must also possess self-learning capabilities. For example, the use of adversarial training enables defence systems to "learn" from attacks, improving resilience. Offensive AI, using methods that adapt to new defence techniques (e.g., developing new attacks based on the analysis of previous results), forces defence to constantly evolve (Shah, 2024). This means that the defence system must quickly react to new attack methods, continuously updating detection models and techniques. Systems that self-learn and use artificial neural networks to analyse large volumes of data can offer high adaptability even to unforeseen types of attacks.

According to the latest studies, the frequency of AI-driven data attacks increased between 2020 and 2025. Specifically, in 2023 the number of phishing attacks rose

by 265% compared to the previous year, many of which were AI-generated. It is projected that by 2025, 45-50% of phishing emails targeting businesses may be generated using AI, with victim response rates rising to 62-65%. Furthermore, in 2024 a rise in AI-based attacks was observed, accounting for approximately 40% of all cyber threats (Martin, 2025). These trends highlighted the need for effective defence strategies against AI-powered attacks. In particular, attention should be paid to multimodal authentication, the use of synthetic data detection algorithms, and the implementation of differential privacy mechanisms. It is also important to restrict open access to model APIs and implement query frequency control to prevent model inversion attacks.

It is critical to note that certain shared vulnerabilities exist across both offensive and defensive systems. These are often linked to data limitations used for training AI models. For example, if the defence system's training dataset is incomplete or inaccurate, the offensive system may find a "breach" by exploiting undocumented vulnerabilities. Conversely, if the offensive system uses adversarial attack methods, its effectiveness may depend on the accuracy and volume of its training data. If these data are insufficiently comprehensive, the offensive system may become vulnerable to countermeasures from the defence side. Therefore, the interaction between offensive and defensive AI systems is complex and multifaceted (Truong *et al.*, 2020). This is a constantly evolving field, where each side's effectiveness depends on the ability to adapt to changing conditions, as well as on real-time self-learning and evolution. In the "AI vs AI" sphere, attacks on information systems using AI take various forms and implementation methods. Some of the most effective approaches include the use of GANs, adversarial input, and reinforcement attack methods. Each of these has distinct features and capabilities for bypassing existing defence systems. GANs are powerful tools for generating fake data or content that can mislead detection systems. A GAN consists of two main components: a generator, which creates new data, and a discriminator, which tries to determine whether the data are fake. These two components operate in a competitive manner: the generator attempts to create data realistic enough to fool the discriminator, while the discriminator tries to distinguish real from generated data (Navidan *et al.*, 2021). In the context of AI attacks, GANs can be used to generate fake images, videos, voice recordings, or texts for social manipulation, phishing, or even manipulation of machine learning outcomes. For instance, GANs can create fake faces so realistic that it is possible to deceive biometric systems.

Adversarial input attacks involve the attacker crafting specially designed inputs that mislead a machine learning model, causing it to make incorrect predictions or decisions. Adversarial inputs are visually imperceptible to humans but can severely disrupt the system's functionality, especially in neural networks that are sensitive to minor input perturbations. This type of attack includes modifying images, text, or other information so that the model misinterprets the data. For example, altering individual pixels on an image may cause a neural network to misclassify it, even though the image appears unchanged to the human eye. This method can be used to bypass security systems based on machine learning, such as facial recognition or suspicious transaction verification systems. Reinforcement attacks use reinforcement learning principles to optimise the attack strategy. In this approach, the offensive agent uses a system of rewards and penalties to build the most effective strategy for evasion. These attacks typically involve interactive learning, where the agent observes the defence system's reactions and adjusts its strategy accordingly. This type of attack is especially dangerous due to its ability to self-optimise. The offensive system learns the defensive AI's behaviour and adapts to its responses, continuously improving its methods to achieve its goal. As a result, reinforcement attacks can effectively bypass even advanced defence algorithms, since the attacking agent can adjust its actions in real time.

Machine learning models used in defence systems can be vulnerable to modification or reconstruction attacks. Model attacks usually aim to gain access to the confidential training data. In particular, hackers may attempt to reconstruct model parameters or even recover data used in training, enabling deception or forgery of the system. Models built on large datasets can be especially vulnerable to this kind of attack, as attackers may exploit partial information about model behaviour to reconstruct internal parameters or training data structure, thereby targeting specific vulnerabilities (He *et al.*, 2020). Context-oriented attacks are another complex method in which the attacking system considers not only the data itself, but also the context in which the data are transmitted. These attacks can target the detection or manipulation of certain aspects of an information system based on the context in which it operates. This allows for more effective attacks, as the system adapts to the specific conditions and vulnerabilities inherent to a given context (Jiang *et al.*, 2022).

Such attacks often leverage a deep understanding of the defence system's internal logic and its responses to certain data types, allowing the attacking agent to tailor its actions to match the conditions of the specific situation in which the attack occurs. Thus, the use of tools such as GANs, adversarial input, and reinforcement attacks enables attackers to bypass security mechanisms efficiently, adapting to changes and improving the strategies in real time. As these methods allow offensive systems to learn and adapt to defences, developing new countermeasures becomes complex but vitally important for ensuring security in the AI era. The integration of AI into cyber threats has created a new class of attacks characterised by adaptability, feedback learning, and high levels of disguise. For defence systems, this means not only increased detection difficulty but also the need for rapid adaptation to environmental changes. Table 1 provided a classification of the main types of AI-based attacks, the descriptions, practical examples, and potential consequences for information security.

***Table 1**. Types of AI attacks and the characteristics*

Type of attacks	Description	Application example	Potential consequences	Attack effectiveness
GAN	Creating fake data that looks realistic to mislead.	Fake images, texts, voices; tricking facial recognition systems.	Reputation damage, fraud, bypassing biometric systems.	High: difficult to distinguish fake data.
Adversarial input	Making minor changes to the input data to manipulate the model's behaviour.	Modifying images or text to deceive the classification system.	False results, incorrect predictions, bypassing security systems.	Medium: depends on the stability of the model.
Reinforcement attack	Adaptive attack that learns with reinforcement to improve efficiency.	Feedback learning to process attack results.	Bypassing complex defence mechanisms, self-optimising attacks.	High: Optimised with feedback.
Model inversion attacks	Recovering training data or internal model parameters based on its behaviour.	Recovering personal data from a machine learning model.	Confidential data leak, breach of confidentiality.	High: especially dangerous when the model is open access.
Context-aware attacks	Attacks that take into account the context of data to manipulate the system under specific conditions.	Data modification to deceive systems that use context (e.g., in recommender systems).	Disruption of recommender systems, interference with the operation of intelligent systems.	Medium-high: effective, but depends on the quality of the context.
Reinforcement learning attack	Attacks using reinforcements to improve efficiency.	Training a model to crack CAPTCHA or bypass Multi-Factor Authentication.	Increased resistance to attacks, bypassing complex authentication mechanisms.	High: 74-87% success rate for CAPTCHA cracking; up to 78% for Multi-Factor Authentication.

***Source:** developed by the author based on A.M. Adawadkar & N. Kulkarni (2022), V. Kumar & D. Sinha (2023), G. Agrawal et al. (2024)*

The comparative analysis of these attacks demonstrated that the most difficult to detect are context-oriented and generative attacks, as these attacks integrate into typical user behaviour or imitate legitimate data. While adversarial input attacks can often be neutralised by increasing model robustness, methods such as reinforcement attack or model inversion require fundamentally new approaches to protection, including blocking external API access, differential privacy enforcement, and regular model parameter updates. Thus, effective counteraction to modern AI attacks requires not only technical upgrades to the defence infrastructure but also a strategic rethinking of security as an adaptive, multi-level process.

Counteraction methods against AI-based attacks: From detection to adaptive defence. Adversarial input attacks are among the most common forms of manipulation in machine learning systems. The main objective of such attacks is to cause the model to produce incorrect results by manipulating input data that appear correct at first glance but may lead to serious errors. Several methods have been developed to combat these attacks. One approach to defence involves designing models capable of detecting anomalous or unauthorised changes in input data. This can be done by comparing inputs to previously known "clean" samples or through additional control models. Another method involves the use of adaptive neural networks that automatically learn from detected adversarial attacks, adjusting the algorithms for increased resistance to similar manipulations. Such approaches enable systems to adapt to new types of attacks without the need for constant retraining. Additionally, specialised architectures may be used, such as input data differentiation or the introduction of noise during processing, which significantly reduces the likelihood of the system misclassifying data due to adversarial changes.

In the context of the dynamic confrontation between offensive and defensive AI systems, each tool has not only advantages but also objective limitations that affect its practical effectiveness. For example, anomaly-based detection systems used to identify unknown attacks demonstrate high sensitivity to non-standard behaviour, but often suffer from high false-positive rates, particularly in complex dynamic environments with high variability of legitimate activity. This significantly complicates the scalability in critical infrastructure. Deep neural networks, used for both offensive and defensive purposes, offer high computational power and self-learning capabilities but are also vulnerable to adversarial examples that can alter model decisions through nearly imperceptible modifications to input data. Moreover, the lack of transparency and explainability complicates security system auditing and certification (Baniecki & Biecek, 2024). Regarding blockchain technologies, often positioned as tools to protect against data tampering and loss of trust in event logs, the key advantage – data immutability – relies on cryptographic block linking and decentralised transaction validation. However, this immutability has a cost: high processing latency, especially in public networks, as well as limited scalability and complexity in integration with traditional information systems. In private blockchains, by contrast, the level of decentralisation decreased, making such systems less secure against node collusion (Gorbenko *et al.*, 2017).

Adaptive defence frameworks based on reinforcement learning demonstrate high effectiveness against evolving threats but require large volumes of training episodes, which are not always compatible with real-world time and resource constraints. Such systems may also fail to adapt correctly to delayed-effect attacks or hidden environmental changes. Multimodal authentication methods, recommended as a defence against generative attacks (including

GAN), enhance resistance to spoofing individual traits but complicate user experience and increase implementation costs due to the need to collect and store multiple types of biometric or behavioural data. Furthermore, these methods are not entirely secure against data fusion level attacks. As a result, no protective mechanism can be considered universal or fully effective in all contexts. For this reason, the modern cybersecurity paradigm increasingly leans towards combined and adaptive approaches that allow the limitations of one technology to be offset by the advantages of another. Determining the right balance between performance, cost, scalability, and protection level remains a key challenge for developers and architects of secure AI systems (Zhang *et al.*, 2020a).

Reinforcement attacks, or adaptive attacks, involve using a reward system to optimise attack strategy. In this case, the attacking system can learn from feedback from previous attacks, allowing it to gradually adapt its strategy for maximum success. This presents a serious security threat, as the attack system can continuously improve, bypassing defensive mechanisms. Adaptive algorithms are used to counter such attacks, which can monitor model behaviour in real time, automatically detecting and neutralising attack strategies (Wylde *et al.*, 2022). Another approach involves limiting the learning strategy of attacking systems, thereby reducing attack effectiveness by restricting the scope for aggressive behaviour optimisation. This includes limiting reward parameters or defining optimisation boundaries.

GANs are becoming increasingly popular not only for creating realistic fake images, videos, and audio, but also for attacks on security systems, particularly biometric ones. GAN-based attacks can lead to serious security breaches, especially when deceiving facial, voice, or fingerprint recognition systems. One method of defending against GAN-based attacks is detecting synthetic data generated by such networks. Specialised models are used to recognise inconsistencies between real and generated data. Additionally, blocking fake data through filtering and verification techniques enables the rejection of suspicious images or other synthetic data types. The integration of blockchain technologies into cybersecurity systems significantly reduces risks related to unauthorised access, data tampering, and malicious impacts on infrastructure. One of blockchain's key advantages is its immutability: each transaction or event is recorded in a decentralised ledger and cannot be altered without the consensus of the majority of network participants. This prevents tampering with security logs or covert editing of system records, which is often the first action taken by an attacker after breaching a system.Furthermore, the decentralised nature of blockchain eliminates the single point of failure typical of centralised storage systems, making DDoS or data hijacking attacks less effective. An attacker would need to simultaneously target numerous nodes to compromise the system's integrity, which is extremely difficult in practice. Blockchain also enables the implementation of more secure access control models via smart contracts. These contracts automatically verify user permissions for performing specific actions, blocking unauthorised attempts to access or modify configurations. As a result, even administrative actions are logged and verified, and access policies cannot be secretly altered.

Another important aspect is the possibility of combining blockchain with AI modules for security event analysis. Since data stored on blockchain cannot be changed, AI models receive guaranteed reliable information for anomaly analysis. This improves the accuracy of threat detection, including abuse, phishing, or credential theft. Additionally, blockchain ensures cryptographic verification of configuration integrity, software updates, and digital certificates. This enables the early detection of attempts to introduce malicious components or firmware changes. Thus, blockchain serves not only as a storage mechanism but also as an active verification and control layer for critical IT infrastructure components. It enhances system trust, supports operational transparency, and ensures resilience to complex targeted attacks. Blockchain is a powerful tool for data security, as it enables the creation of immutable records that cannot be altered after entry into the system. This is particularly useful in preventing attacks aimed at altering or falsifying data. Using blockchain for data security includes secure authentication and verification, allowing data to be stored in immutable form and ensuring its authenticity. Moreover, blockchain can be applied to create distributed ledgers where all transactions are recorded in multiple locations, making data manipulation impossible without access to the entire network (Saleh, 2024).

In the context of a rising number of attacks aimed at leaking confidential information, data protection is becoming a crucial component of cybersecurity. One of the most effective protection methods is the use of differential privacy techniques, which enable data processing without compromising confidentiality. This method allows machine learning models to work with data without disclosing private information (Lapid *et al.*, 2024). It adds noise to data to prevent the identification of individuals or objects. An important element of protection is also data anonymisation before it enters the system, preserving its utility for training while preventing privacy breaches. These AI-focused defence methods are just one part of a broader strategy aimed at enhancing security in the modern digital environment. Going forward, with the advancement of AI and new technologies, it is crucial to continually update protection approaches to remain prepared for emerging cybersecurity challenges.

In the context of increasing cyber threats leveraging AI capabilities, the development of effective defence mechanisms has become critically important. Methods for countering AI-based attacks are becoming increasingly complex, combining both classical security approaches and innovative solutions based on self-learning, anomaly detection, and secure data processing. Table 2 presented the main methods of protection against AI-based attacks, including descriptions of the characteristics, advantages, disadvantages, and areas of application.

Table 2. *Key methods for protecting against AI-based attacks*

Method name	Principle of operation	Advantages	Limitations/ disadvantages	Areas of application
Adversarial detection	Analysing incoming data for suspicious changes.	Improves classification accuracy; adapts to new attacks.	May give false positive results; requires additional resources.	Autonomous systems, biometrics, Internet of Things.
Adaptive training	The model learns to recognise new types of attacks during operation.	Increases resistance to unpredictable attacks.	Requires constant monitoring and computing power.	Critical infrastructures, financial systems.
Protection against GAN attacks	Detection of synthetic data created by generative networks.	Protection against fake images, videos, biometric forgeries.	High requirements for the accuracy of detection models.	Biometric systems, digital security.
Noise input/ differential privacy	Adding statistical noise to data to preserve privacy.	Provides privacy without losing functionality.	May reduce model accuracy.	Healthcare, user analytics, government services.
Blockchain protection	Recording transactions in immutable distributed ledgers.	Immutability; transparency; resistance to counterfeiting.	Scalability limitations; high integration cost.	Data protection, audit, document management.
Data anonymisation	Removing or masking personal information before analysis.	Reducing the risk of data leakage.	Loss of accuracy and usefulness in certain scenarios.	Educational platforms, medical research.
Feedback limitation for reinforcement attacks	Reducing the information an attacker can obtain for training.	Reduces the effectiveness of adaptive attacks.	Can slow down the system; difficult to configure.	Industrial systems, autonomous devices.

Source: *developed by the author based on A. Aldahdooh et al. (2022), Y. Liu et al. (2022), C. Barreto et al. (2023)*

As shown in Table 2, each of the methods had its own unique properties and limitations, which determined the need for combination into multi-level security systems. None of the approaches provided absolute protection against all types of attacks, but together these approaches significantly increased the security level of AI-based systems. Further research had to be focused on integrating these solutions, adapting to new threats, and developing dynamic models capable of independently detecting and neutralising even previously unknown types of attacks. To ensure resilience against AI-powered attacks, developers had to consider a number of recommendations adapted to the specifics of each type of threat. In particular, to protect against attacks using GAN, it was advisable to implement multimodal authentication that combined several levels of verification – biometric, behavioural, and traditional (passwords, tokens). This made it more difficult to use fake images, voices, or videos to bypass security systems. It was also recommended to apply algorithms for detecting synthetic data, including deepfake detectors trained to distinguish signs of artificial generation.

Against adversarial input attacks – that is, the introduction of imperceptible changes into input data to manipulate model behaviour – adversarial training proved effective, where the model was also trained on noisy or distorted examples. Additionally, it was worth implementing normalisation mechanisms and permissible range checks to filter suspicious input signals before processing. To counter reinforcement attacks, i.e., adaptive attacks that learned through feedback, it was important to implement dynamic and variable authentication policies so that it would be difficult for the attacking model to develop a stable strategy. Multifactor authentication (including the use of temporary codes, biometrics, and contextual information such as geolocation) significantly complicated the learning process of attacking systems.

Regarding model inversion attacks, where confidential training data were reconstructed based on the model's output results, it was recommended to introduce differential privacy mechanisms that added controlled noise to the outputs and reduced the risk of leakage. It was also necessary to limit open access to the model's API, implement query frequency control, and use authorisation layers. In the case of context-oriented attacks, which used manipulations with environment or user variables, it was advisable to perform consistency checks (e.g., verifying whether geolocation matched the user's behavioural pattern) as well as to verify data sources and timestamps. Anti-fraud modules detecting anomalous behaviour were worth applying.

In general, one of the effective approaches was the use of blockchain technologies for storing critical access logs and user actions. Due to its immutability – ensured by cryptographic hashing and consensus algorithms – blockchain enabled the recording of all events in such a way that no edits could be made by an attacker without detection. This not only increased transparency and control but also made it more difficult to cover traces of an attack or alter critical system parameters. It was also recommended to carry out regular system testing for resistance to attacks (at least quarterly) using generative methods and to implement a Zero Trust architecture, which assumed no trust in system components without verification, even within the internal perimeter. This enabled timely detection and isolation of potentially dangerous activity even at early stages.

Classification of AI attacks and countermeasures in the dynamics of evolutionary confrontation. The analysis of the effectiveness of offensive AI models and the response speed of defence systems revealed a significant shift in the balance of power in the modern cyberspace. In particular, AI integrated into offensive mechanisms demonstrated

high efficiency in bypassing classical threat detection tools. For example, the use of GANs to generate phishing emails or malicious traffic enabled attack success rates of 87-93%, significantly complicating the detection by signature-based systems. Offensive models with reinforcement learning, used for breaking CAPTCHAs or bypassing multifactor authentication, achieved an attack success rate of approximately 78%. In turn, defence systems based on traditional machine learning algorithms (decision trees, Support Vector Machine) showed an average intrusion response time within 4.2-6.5 seconds. More effective were deep neural networks, particularly Long Short-Term Memory or transformers, which allowed anomalies in traffic to be detected within 2.1-3.4 seconds. The best results were demonstrated by hybrid approaches that combined behavioural analysis and anomaly detection models, with a response time of around 1.7 seconds, although with a higher false-positive rate – up to 9%. Additionally, in evolutionary confrontation scenarios – where offensive and defensive AI systems continuously adapted to each other – there was a recorded 15-20% reduction in attack effectiveness after five iterations of self-learning by the defence model, indicating the promise of an adaptive approach to cyber defence (Hassija *et al.*, 2019; Zhang *et al.*, 2020b).

In the context of the rapid development of offensive AI systems, studying the real-world impact on existing cyber defence measures is becoming increasingly relevant. Despite rising investment in intelligent defence solutions, practical examples show that attackers are already actively using generative models, training attacks, and other forms of adaptive AI to bypass even complex defence mechanisms. To gain a deeper understanding of the scale of this threat, it is useful to examine specific cases in which AI-based attacks have demonstrated high effectiveness in real or experimental conditions. Real-world cases increasingly show the use of AI in offensive scenarios, particularly through generative models capable of bypassing modern defence mechanisms. A notable example was a study where GANs were used to create artificial facial images that successfully bypassed biometric authentication systems based on FaceNet. In this experiment, the generated images showed up to 85% likelihood of misidentification, allowing attackers to access accounts without physical presence. The researcher found that even models trained on large datasets such as Labeled Faces in the Wild could not distinguish a GAN-generated fake face from a real one. Another example was a study in which the researcher used specially printed glasses with textures generated by AI to trick the facial recognition system of Face++, simulating another person's appearance. This attack enabled targeted identity imitation (i.e., appearing as a specific individual), making the method potentially dangerous for banking or access control systems.

In the field of text-based system hacking, the author demonstrated how GAN-generated phishing emails, stylistically adapted to the victim's correspondence, managed to bypass more than 70% of spam filters, including those using deep learning models on mail server sides. The click-through rate on malicious links reached 32%, far exceeding the average figures for classical phishing campaigns (10-12%). Another case was a study in which it was proven that offensive models with data inversion functionality, using access to the API of a protected model, were able to reconstruct facial images on which the model had been trained. This poses significant risks of personal data leakage, even without direct access to the data.

Defence methods used to counter AI-based attacks are based on various principles, each with its own strengths and limitations depending on the type of threat. For example, blockchain technology relies on a distributed ledger in which all data are recorded in the form of blocks linked together by cryptographic hashes. Each new block contains the hash of the previous one, making it impossible to alter previously entered information without breaking the entire chain. This ensures data "immutability" – if an attacker attempts to change even a single record, all subsequent blocks must also be changed and synchronised with the majority of network nodes, which is virtually impossible in practice. This mechanism effectively protects systems from transaction forgery, falsification of security logs, or data poisoning in AI model training processes.

Given the growing cyber threats and constant improvement of offensive AI models, a comparative analysis of key defence methods was timely, allowing the identification of the strengths and weaknesses. This analysis considered not only the level of confidentiality and resilience to attacks but also the impact on model accuracy, resource intensity, and processing time. This enabled a reasoned selection of the most appropriate approaches depending on application specifics, available computational resources, and risk levels. The summarised results of this analysis are presented in Table 3, which showed a comparison of the effectiveness of methods for ensuring AI system confidentiality and resilience to attacks.

Table 3.*Comparison of the effectiveness of methods for ensuring confidentiality and resilience of AI systems to attacks*

Method	Privacy level	Resistance to attacks	Impact on model accuracy	Resource intensity	Processing time	Application features
Federated learning	High	Medium	Minimum (up to -3%)	High	Moderate	Local learning, reducing the risk of leakage; vulnerability to model poisoning.
Homomorphic encryption	Very high	High	Minor	Very high	Very slow	Computation on encrypted data; efficient in cloud services.
Differential privacy	High	Medium-high	Loss of accuracy up to 5-7%	Medium	Moderate	Protecting personal data by adding noise; used in fintech, eHealth.

Table 3. Continued

Method	Privacy level	Resistance to attacks	Impact on model accuracy	Resource intensity	Processing time	Application features
Blockchain	High	High	Does not affect	High	High	Transparency, immutability, audit; access control, logging.
Adversarial training	Low	High (up to +25%)	Loss of accuracy 2-4%	Medium-high	Medium	Increases resistance to attacks by training on modified data.
Combined approaches	Very high	Very high (up to +40%)	Moderate loss of accuracy (3-6%)	High-very high	Moderate	Integration of various mechanisms: FL+DP+Blockchain; maximum adaptability.

Source: *developed by the author*

As shown in Table 3, none of the methods is universal, and the choice of an optimal approach should depend on specific objectives, the threat context, and technical limitations. For example, homomorphic encryption ensures the highest level of confidentiality but is extremely resource-intensive and slow. While adversarial training significantly increases resistance to attacks, it does not protect against data leakage. Particularly promising are combined models that integrate the advantages of different approaches. For instance, the integration of blockchain and differential privacy enables the creation of systems that are simultaneously private, transparent, and resilient to attacks. These integrated strategies demonstrate the best performance under conditions of dynamic development of offensive models and growing cybersecurity challenges.

In the case of Disney, a hacker distributed malware disguised as an AI art application via platforms such as GitHub. This enabled unauthorised access to an employee's computer, resulting in the theft of approximately 1.1 TB of confidential information, including Slack messages, client data, and internal corporate discussions (Mojadad, 2025). British retailers fell victim to cyberattacks, allegedly carried out by the Scattered Spider group, which uses phishing and SIM swapping to penetrate networks. These attacks disrupted the operation of online services and internal company systems (The Times, 2025). Meanwhile, the CEO of advertising giant Wire and Plastic Products was targeted by fraudsters who used a deepfake of the voice to arrange a fake video meeting with the company's management, attempting to obtain confidential information (BizzComunity, 2024). These cases underline the need to implement multi-layered protection strategies, including multimodal authentication, synthetic data detection algorithms, and mechanisms for differential privacy. It is also essential to restrict open access to model APIs and introduce query rate control to prevent model inversion attacks.

An important element in analysing the effectiveness of modern offensive and defensive AI systems is studying the dynamics of real cyber incidents across individual companies. Particularly illustrative is the case of the Twitter (X) platform, which in the latest years, has repeatedly become a target of coordinated cyberattacks involving social engineering, API abuse, and generative phishing schemes. Table 4 presented summarised statistics for the 2020-2025 period, demonstrating the escalation of threats and the shortcomings of classical defence approaches.

Table 4*. Dynamics of cyberattacks on Twitter (X) in 2020-2025*

Year	Number of attacks / affected accounts	Growth (%)
2020	≈130 accounts hacked (known accounts of politicians, companies, crypto exchanges)	-
2021	Isolated attack attempts, no significant mass breaches recorded	-98%
2022	5.4 million accounts (leaked due to API vulnerability)	+4,050%
2023	≈200 million accounts (mass leak from forum)	+3,600%
2024	≈249 million accounts (leaked due to reuse of old data and new phishing campaigns)	+24.5%
2025	Large-scale DDoS attack causes global outage of X platform; no data leaks reported	Data missing

Source: *developed by the author based on J. Tidy & D. Molloy (2020), L. Abrams (2022), A. Mascellino (2023), Kaaviya (2024), M. Chapman & B. Ortutay (2025)*

The analysis of Table 4 revealed a clear trend: following a large-scale attack in 2020, the company managed to temporarily reduce malicious activity, but from 2022 a sharp increase in attacks has been observed, mainly due to technological vulnerabilities and new threat vectors. Particular attention should be paid to 2022, when a critical vulnerability in the Twitter API was discovered, allowing attackers to mass collect users' personal data. In the period 2023-2025, the nature of attacks changed: whereas social engineering previously dominated, the key threats now include AI-based automated tools, generative phishing platforms, and bot networks. This highlights the need not only for rapid updates to technical protection tools, but also for a rethinking of security architecture based on Zero Trust principles, dynamic authentication, behavioural analysis, and continuous self-learning of defence systems. As of 2025, no official information has been recorded regarding large-scale personal data breaches from the Twitter (X) platform. Despite reports of a DDoS attack and temporary service disruptions, data on the total number of affected accounts or a rise in incidents is currently unavailable or unconfirmed by open sources. Therefore, the assessment of cyberattack dynamics for 2025 has not yet been presented.

Discussion

The discussion of research findings regarding the use of AI in the context of cybersecurity, particularly the "AI vs AI" confrontation, revealed several key aspects that deserve more detailed analysis. The examined types of attacks, methods of the implementation, and the interaction of offensive and defensive AI systems have a significant impact on the current situation in the field of cyber defence. It is important to note the classification of offensive AI by various criteria. Such differentiation allows for a more detailed study of the strategies of AI-driven attacks. Each type of attack, whether information-analytical, generative, or aggressively influential, requires specific protection approaches. For example, information-analytical agents are aimed at collecting and processing data to identify vulnerabilities, which may pose a threat to any system if effective monitoring and behavioural analysis mechanisms are not applied. Generative models, particularly GANs, enable the creation of fake data, which significantly complicates the detection of such attacks. These models are already actively used in phishing campaigns, requiring increased attention to the development of content authenticity verification tools. Aggressively influential models, which affect system functionality, especially through the introduction of malicious data into training datasets or input data manipulation, represent a serious threat to system stability. Q. Pan *et al.* (2022), in the study, proposed a classification of offensive AI by type of impact – information-analytical, generative, and aggressively influential – emphasising that each group requires specific approaches to detection and neutralisation. The alignment with current results lay in recognising the high threat posed by generative models, particularly the use in phishing campaigns, as well as the danger of aggressively influential AI that manipulates training data. At the same time, differences lie in protection approaches: the authors focused on the classification and characteristics of threats, while the presented model emphasised the integration of protection mechanisms, including blockchain registration and behavioural anomaly verification in critical infrastructure systems.

The classification by learning method also revealed that pre-trained models were more effective under conditions of mass attacks, whereas adaptive systems proved significantly more dangerous due to the ability to change behaviour in real time, adjusting to the responses of the protection system. This highlights the importance of using self-learning systems in countering attacks, especially through the introduction of new learning methods in protection systems. R. Sabitha *et al.* (2023), in the study, classified offensive AI by learning method, finding that pre-trained models demonstrated higher efficiency in executing mass attacks due to optimised structure and execution speed. At the same time, the authors gave particular attention to adaptive systems, which the scientists considered much more dangerous due to the ability to change behaviour in real time, responding to defence actions. H.I. Kure *et al.* (2022) emphasised the need to implement self-learning defence solutions capable of mimicking or outpacing the evolution of attacks, which partially aligns with the current conclusions regarding the integration of dynamic, flexible algorithms into the protection of critical infrastructure. However, unlike the authors, the current model focused not only on adaptation but also on action transparency through blockchain mechanisms.

An important aspect is the real-time interaction between offensive and defensive AI. During such confrontations, offensive AI is able to adapt its strategies by seeking new vulnerabilities in defence systems, which requires continuous improvement of protection models. The cyclical nature of this struggle, when offensive systems change the strategies and defensive systems adapt to new threats, increases the complexity of developing effective protection systems. In turn, the use of methods such as adversarial training allows defensive systems to improve the resilience based on learning from attacks. R.R. Kethireddy (2023), in the study, focused on the dynamic interaction between offensive and defensive AI in real time. The author noted that the ability of offensive AI to adapt to the behaviour of defence systems creates a constant challenge for defensive models, which must not only respond but also anticipate possible attack strategies. The author also emphasised the cyclical nature of this process – with changing attack approaches comes an increased need for the evolution of defence mechanisms. This resonates with the current study, which also highlighted the necessity of continuous protection updates. However, unlike the author, the current study focused on integrating blockchain technologies as a means of ensuring data integrity in this dynamic interaction, rather than solely on adversarial training.

The adaptability of offensive and defensive systems, as noted, is a key component of effective confrontation. During the study, it was found that offensive AI, using self-learning methods, can quickly adjust its strategy in real time, making the battle between attacking and defensive systems even more dynamic. This requires defensive systems to make considerable efforts to timely detect new attack methods and adapt to such methods. At the same time, as noted by V. Nesterov (2023), the integration of AI into data engineering enables advanced automation and predictive analysis, which contributes to faster response and informed decision-making in complex information environments. The implementation of common standards and improved algorithms enhances the efficiency and security of data processing, creating a more stable foundation for the operation of adaptive defence systems. M.T. Hossain *et al.* (2024), in the study, concluded that despite the high adaptability of offensive systems, defensive AI models should not focus on instant reaction but rather on building a resilient architecture capable of withstanding a wide range of attacks without the need for constant retuning. The authors believed that excessive adaptability of defence systems could lead to increased instability in the operation. This contradicts the current results, which proved that it is precisely dynamic adaptation and the integration

of self-learning mechanisms that are key to effectively detecting and neutralising evolving attacks. Thus, there is a conceptual divergence between the authors' approaches and the current study regarding the appropriateness of active adaptation in defence.

The issue of the evolution of offensive and defensive systems is of particular importance. According to the study results, both systems can evolve in real time, allowing for improved efficiency in the process of interaction. Attacks using GANs proved to be among the most dangerous due to the ability to generate fake data that can be used to bypass protection. At the same time, defence systems, using advanced anomaly detection methods, can partially counter such attacks, but this requires constant updating and improvement of models. C. Yinka-Banjo & O.A. Ugot (2020), in the work, argued that although GAN-based attacks demonstrate a high level of complexity, these attacks are not the defining threat factor. In the scientists' view, combined attacks that integrate social engineering with AI technologies are more dangerous as these attacks affect not only technical but also human elements of the system. This differs from the current findings, where GAN-oriented attacks were recognised as the most dangerous due to the ability to generate convincing fake data that is difficult to detect even by updated systems. Thus, the focus of the authors' and current studies diverges both in identifying the main threat and in approaches to its neutralisation.

C. Zhang *et al.* (2023), in the work, emphasised that offensive systems, especially those using GANs, were among the most dangerous due to the ability to generate fake data that can deceive defence systems. In addition, the authors drew attention to the evolution of offensive strategies, where systems adapt to new conditions in real time, posing the challenge of continuous algorithmic improvement for defensive models. To ensure effective protection, the authors proposed the use of anomaly detection methods and the integration of self-learning systems into defence strategies. Compared to the current study, there are shared points regarding the need for system adaptability and the use of generative models as threats. However, the current work paid more attention to learning methods in the context of attacks and defence, whereas the authors focused on the evolution of offensive strategies and general approaches to improving protection.

Despite all the advantages of attacking models, it is important to note that there are also certain weaknesses in both systems. For example, attacks using adversarial input may only be effective if the input data are properly prepared to bypass defence systems. At the same time, data limitations for model training are a potential vulnerability for offensive systems, which may fail due to insufficient coverage of all possible attack types. A. Chakraborty *et al.* (2021), in the current study, reached similar conclusions, emphasising that the effectiveness of attacks using adversarial input largely depends on the accuracy of input data preparation. The authors also noted that offensive models are vulnerable when there is limited access to high-quality and representative training datasets. This confirmed the results of the current study, which found that flaws or insufficiencies in data can significantly reduce attack success, just as the effectiveness of protection depends on the ability to identify the specifics of prepared attacks.

M. Aminu *et al.* (2024), in the study, focused on the importance of developing specialised algorithms for adapting offensive systems to new threats. In particular, the authors proposed the integration of deep learning methods for implementing more complex attacks that can effectively bypass traditional defence mechanisms. The authors also stressed the need for continuous updating and development of defence strategies based on flexibility and self-learning principles, to be able to respond quickly to new types of attacks, especially in complex dynamic environments. Comparing these results with the current ones, it is clear that both approaches agree on the importance of adaptability and the evolution of offensive and defensive systems. However, the current emphasis on generative models and the ability to produce fake data does not fully align with the authors' approach, which focused on the use of deep learning methods for attacks. The current work devoted more attention to real-time interaction and adaptation, whereas the authors concentrated on the need for ongoing model updates to counter new types of attacks.

Thus, the results of the study point to the importance of developing new approaches to creating both offensive and defensive AI systems. Success in the "AI vs AI" confrontation depends on both sides' ability to rapidly adapt to changing conditions and improve the strategies. These studies open new opportunities for ensuring security in the AI era, but at the same time pose new challenges for defence systems. H. Rauf *et al.* (2025) also highlighted the importance of adaptation in the confrontation between offensive and defensive AI systems. However, in the work, the focus was more on technological innovations in the field of attacks, such as the use of new generative model methods to bypass protection. At the same time, the approach focused on the need to invest in new defence technologies capable of overcoming these new attack strategies. Meanwhile, the current results focused more on the dynamic adaptation of both sides, where the key aspect was not only technological progress but also the speed of response to changing conditions. Despite some differences in approach, it can be noted that both studies agree on the necessity of rapid adaptation and improvement of strategies for effective confrontation in the AI era.

The reviewed studies collectively highlighted the growing complexity and dynamism of AI-driven confrontations in cybersecurity. Despite differences in focus – ranging from attack typologies and learning models to defence mechanisms and system adaptability – a common thread was the critical importance of real-time response, continuous model evolution, and integration of advanced technologies such as generative models, self-learning algorithms, and blockchain. The current findings emphasised that the effectiveness of both offensive and defensive AI systems hinges not only on technical sophistication but also on

their capacity for rapid adaptation to emerging threats. This positions the "AI vs AI" paradigm as a defining challenge for the future of cyber defence.

Conclusions

This study analysed the use of AI both for launching attacks on information systems and for the protection within the context of the evolutionary confrontation "AI vs AI". A detailed classification of offensive models was carried out according to the type of action (information-analytical, generative, aggressively influential), training approaches (static, continuous, reinforcement), target of the attacks (data, models, users), and level of autonomy (human-oriented, semi-autonomous, fully autonomous). This approach made it possible to identify the riskiest attack configurations, particularly generative models that adapt in real time and are aimed at bypassing deep learning defence systems. The dynamics of interaction between offensive and defensive AI systems in the context of evolutionary confrontation were analysed, where both sides demonstrated the ability to learn, adapt, and develop counter-strategies in real time. The classification of attacks included information-analytical (gathering and analysing confidential information to build targeted attacks), generative (creating fake data or manipulative samples capable of bypassing classifiers), and aggressively influential (altering the behaviour of users or the system itself). The division by level of autonomy (from human-oriented to fully autonomous agents) made it possible to identify the most dangerous configurations – autonomous generative systems that learn during the attack process. During modelling, it was found that agents with reinforcement learning demonstrated a high level of penetration into protected systems even in the presence of traditional security mechanisms. For example, generative attacks produced up to 85% false-positive results in biometric systems based on FaceNet. Defensive strategies, in turn, showed varying levels of effectiveness: adversarial training increased resilience to attacks by 25%, but depended on the completeness of the training sample; the use of blockchain ensured high reliability and access control, but affected system performance in real-time mode. The most balanced results were achieved through combined approaches – the integration of FL, DP, blockchain, and multimodal authentication made it possible to achieve up to +40% resilience with an acceptable accuracy loss (3-6%). This emphasised the importance of multi-component, adaptive, and context-sensitive solutions in countering high-level intelligent threats. Further research should focus on reducing the resource intensity of defence systems, developing effective methods for detecting contextual attacks and generative distortions, and expanding training datasets to improve the resilience of models to new types of threats.

Acknowledgements

None.

Funding

None.

Conflict of Interest

None.

References

[1] Abrams, L. (2022). Twitter confirms zero-day used to expose data of 5.4 million accounts. *Bleeping Computer*. Retrieved from https://surl.li/govnio.

[2] Adawadkar, A.M., & Kulkarni, N. (2022). Cyber-security and reinforcement learning – a brief survey. *Engineering Applications of Artificial Intelligence*, 114, article number 105116. doi: 10.1016/j.engappai.2022.105116.

[3] Agrawal, G., Kaur, A., & Myneni, S. (2024). A review of generative models in generating synthetic attack data for cybersecurity. *Electronics*, 13(2), article number 322. doi: 10.3390/electronics13020322.

[4] Al-Ahmadi, S., Alotaibi, A., & Alsaleh, O. (2022). PDGAN: Phishing detection with generative adversarial networks. *IEEE Access*, 10, 42459-42468. doi: 10.1109/ACCESS.2022.3168235.

[5] Aldahdooh, A., Hamidouche, W., Fezza, S.A., & Déforges, O. (2022). Adversarial example detection for DNN models: A review and experimental comparison. *Artificial Intelligence Review*, 55, 4403-4462. doi: 10.1007/s10462-021-10125-w.

[6] Aminu, M., Akinsanya, A., Dako, D.A., & Oyedokun, O. (2024). Enhancing cyber threat detection through real-time threat intelligence and adaptive defense mechanisms. *International Journal of Computer Applications Technology and Research*, 13(8), 11-27. doi: 10.7753/IJCATR1308.1002.

[7] Baniecki, H., & Biecek, P. (2024). Adversarial attacks and defenses in explainable artificial intelligence: A survey. *Information Fusion*, 107, article number 102303. doi: 10.1016/j.inffus.2024.102303.

[8] Barreto, C., Reinert, O., Wiesinger, T., & Franke, U. (2023). Duopoly insurers' incentives for data quality under a mandatory cyber data sharing regime. *Computers & Security*, 131, article number 103292. doi: 10.1016/j.cose.2023.103292.

[9] BizzComunity. (2024). *WPP CEO Mark Read targeted by deep fake AI scam*. Retrieved from https://www.bizcommunity.com/article/wpp-ceo-mark-read-targeted-by-deep-fake-ai-scam-340068a.

[10] Calvaresi, D., Dicente Cid, Y., Marinoni, M., Dragoni, A.F., Najjar, A., & Schumacher, M. (2021). Real-time multi-agent systems: Rationality, formal model, and empirical results. *Autonomous Agents and Multi-Agent Systems*, 35, article number 12. doi: 10.1007/s10458-020-09492-5.

[11] Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., & Mukhopadhyay, D. (2021). A survey on adversarial attacks and defences. *CAAI Transactions on Intelligence Technology*, 6(1), 25-45. doi: 10.1049/cit2.12028.
[12] Chapman, M., & Ortutay, B. (2025). Elon Musk claims X being targeted in "massive cyberattack" as service goes down. *The Associated Press*. Retrieved from https://apnews.com/article/x-musk-twitter-outage-california-0268a8b035aaa277c0287e7c82b6081e.
[13] Gopireddy, R.R. (2024). Securing AI systems: Protecting against adversarial attacks and data poisoning. *Journal of Scientific and Engineering Research*, 11(5), 276-281.
[14] Gorbenko, I., Nariezhnii, O., & Kudryashov, I. (2017). Construction method and features of one class of cryptographic discrete signals. In *Proceedings of the 4th international scientific-practical conference problems of infocommunications. Science and technology* (pp. 156-160). Kharkiv: IEEE. doi: 10.1109/INFOCOMMST.2017.8246371.
[15] Hasan, M. (2024). A study on the integration of blockchain technology for enhancing data integrity in cyber defense systems. *Journal of Digital Transformation, Cyber Resilience, and Infrastructure Security*, 8(12), 21-30.
[16] Hassija, V., Chamola, V., Saxena, V., Jain, D., Goyal, P., & Sikdar, B. (2019). A survey on IoT security: Application areas, security threats, and solution architectures. *IEEE Access*, 7, 82721-82743. doi: 10.1109/ACCESS.2019.2924045.
[17] He, Y., Meng, G., Chen, K., Hu, X., & He, J. (2020). Towards security threats of deep learning systems: A survey. *IEEE Transactions on Software Engineering*, 48(5), 1743-1770. doi: 10.1109/TSE.2020.3034721.
[18] Hossain, M.T., Afrin, R., & Biswas, M.A. (2024). A review on attacks against artificial intelligence (AI) and their defence image recognition and generation machine learning, artificial intelligence. *Control Systems and Optimization Letters*, 2(1), 52-59. doi: 10.59247/csol.v2i1.73.
[19] Jiang, Y., Wu, S., Yang, H., Luo, H., Chen, Z., Yin, S., & Kaynak, O. (2022). Secure data transmission and trustworthiness judgement approaches against cyber-physical attacks in an integrated data-driven framework. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 52(12), 7799-7809. doi: 10.1109/TSMC.2022.3164024.
[20] Kaaviya. (2024). Massive 9.4GB Twitter data leaked online – 200 million records exposed. *Cyber Security News*. Retrieved from https://cybersecuritynews.com/massive-9-4gb-twitter-data-leaked-online/.
[21] Karl, L., & Potter, K. (2023). *Knowledge transfer in machine learning, adaptive learning with pretrained models*. doi: 10.31219/osf.io/dy5v7.
[22] Kethireddy, R.R. (2023). AI-augmented threat response systems with real-time adaptive defense. *International Journal of Artificial Intelligence Research and Development,* 1(1), 62-71.
[23] Kumar, S., Dwivedi, M., Kumar, M., & Gill, S.S. (2024). A comprehensive review of vulnerabilities and AI-enabled defense against DDoS attacks for securing cloud services. *Computer Science Review*, 53, article number 100661. doi: 10.1016/j.cosrev.2024.100661.
[24] Kumar, V., & Sinha, D. (2023). Synthetic attack data generation model applying generative adversarial network for intrusion detection. *Computers & Security*, 125, article number 103054. doi: 10.1016/j.cose.2022.103054.
[25] Kure, H.I., Islam, S., & Mouratidis, H. (2022). An integrated cyber security risk management framework and risk predication for the critical infrastructure protection. *Neural Computing and Applications*, 34, 15241-15271. doi: 10.1007/s00521-022-06959-2.
[26] Lapid, R., Dubin, A., & Sipper, M. (2024). Fortify the guardian, not the treasure: Resilient adversarial detectors. *Mathematics*, 12(22), article number 3451. doi: 10.3390/math12223451.
[27] Liu, Y., Li, Y., Deng, G., Liu, Y., Wan, R., Wu, R., Ji, D., Xu, S., & Bao, M. (2022). Morest: Model-based RESTful API testing with execution feedback. In M.B. Dwyer (Ed.), *Proceedings of the 44th international conference on software engineering* (pp. 1406-1417). New York: Association for Computing Machinery. doi: 10.1145/3510003.3510133.
[28] Marino, D.L., Wickramasinghe, C.S., Rieger, C., & Manic, M. (2025). Self-supervised and interpretable anomaly detection using network transformers. *IEEE Transactions on Industrial Informatics*, 21(5), 4252-4261. doi: 10.1109/TII.2025.3534443.
[29] Martin, K. (2025). AI Cyber Attack Statistics 2025. *TechAdvisor*. Retrieved from https://surl.li/whsnri.
[30] Mascellino, A. (2023). *Over 200 million twitter users' details leaked on hacker forum*. Retrieved from https://www.infosecurity-magazine.com/news/over-200m-twitter-users-details/.
[31] Mojadad, I. (2025). Calif. man hacked Disney worker's computer, stole vast store of company data. *Infosecurity Magazine*. Retrieved from https://www.sfgate.com/disneyland/article/disney-hack-stolen-data-20307312.php.
[32] Muhuri, P.S., Chatterjee, P., Yuan, X., Roy, K., & Esterline, A. (2020). Using a long short-term memory recurrent neural network (LSTM-RNN) to classify network attacks. *Information*, 11(5), article number 243. doi: 10.3390/info11050243.
[33] Navidan, H., Moshiri, P.F., Nabati, M., Shahbazian, R., Ghorashi, S.A., Shah-Mansouri, V., & Windridge, D. (2021). Generative adversarial networks (GANs) in networking: A comprehensive survey & evaluation. *Computer Networks*, 194, article number 108149. doi: 10.1016/j.comnet.2021.108149.
[34] Nesterov, V. (2023). Integration of artificial intelligence technologies in data engineering: Challenges and prospects in the modern information environment. *Bulletin of Cherkasy State Technological University*, 28(4), 82-92. doi: 10.62660/2306-4412.4.2023.82-90.

[35] Pan, Q., Wu, J., Bashir, A.K., Li, J., & Wu, J. (2022). Side-channel fuzzy analysis-based AI model extraction attack with information-theoretic perspective in intelligent IoT. *IEEE Transactions on Fuzzy Systems*, 30(11), 4642-4656. doi: 10.1109/TFUZZ.2022.3172991.

[36] Pu, G., Wang, L., Shen, J., & Dong, F. (2020). A hybrid unsupervised clustering-based anomaly detection method. *Tsinghua Science and Technology*, 26(2), 146-153. doi: 10.26599/TST.2019.9010051.

[37] Ragmani, A., Elomri, A., Abghour, N., Moussaid, K., Rida, M., & Badidi, E. (2020). Adaptive fault-tolerant model for improving cloud computing performance using artificial neural network. *Procedia Computer Science*, 170, 929-934. doi: 10.1016/j.procs.2020.03.106.

[38] Rauf, H., Shah, S.I., Ali, T., Gul, H., & Soomro, M. (2025). Using generative AI for simulating cyber security attacks and defense mechanisms: A new approach to ai-driven cyber threat modeling. *Spectrum of Engineering Sciences*, 3(3), 361-381.

[39] Sabitha, R., Gopikrishnan, S., Bejoy, B.J., Anusuya, V., & Saravanan, V. (2023). Network based detection of IoT attack using AIS-IDS model. *Wireless Personal Communications*, 128, 1543-1566. doi: 10.1007/s11277-022-10009-4.

[40] Saleh, A.M. (2024). Blockchain for secure and decentralized artificial intelligence in cybersecurity: A comprehensive review. *Blockchain: Research and Applications*, 5(3), article number 100193. doi: 10.1016/j.bcra.2024.100193.

[41] Santhi, K., Shri, M.L., Joshi, S., & Sharma, G. (2024). AI in defence and ethical concerns. In *Proceedings of the second international conference on emerging trends in information technology and engineering* (pp. 1-7). Vellore: IEEE. doi: 10.1109/ic-ETITE58242.2024.10493592.

[42] Shah, B. (2024). Adaptive defense strategies for protecting AI models from evasion attacks in adversarial machine learning. *Aitoz Multidisciplinary Review*, 3(1), 323-337.

[43] The Times. (2025). *Who are Scattered Spider hackers linked to the M&S cyberattack?* Retrieved from https://www.thetimes.com/uk/technology-uk/article/scattered-spider-hackers-ms-cyber-attack-lr3w92kcl.

[44] Tidy, J., & Molloy, D. (2020). Twitter hack: 130 accounts targeted in attack. *BBC*. Retrieved from https://www.bbc.com/news/technology-53445090.

[45] Truong, T.C., Diep, Q.B., & Zelinka, I. (2020). Artificial intelligence in the cyber domain: Offense and defense. *Symmetry*, 12(3), article number 410. doi: 10.3390/sym12030410.

[46] Wu, J., Li, L., Shi, F., Zhao, P., & Li, B. (2022). A two-stage power system frequency security multi-level early warning model with DS evidence theory as a combination strategy. *International Journal of Electrical Power & Energy Systems*, 143, article number 108372. doi: 10.1016/j.ijepes.2022.108372.

[47] Wylde, V., Rawindaran, N., Lawrence, J., Balasubramanian, R., Prakash, E., Jayal, A., Khan, I., Hewage, C., & Platts, J. (2022). Cybersecurity, data privacy and blockchain: A review. *SN Computer Science*, 3, article number 127. doi: 10.1007/s42979-022-01020-4.

[48] Yinka-Banjo, C., & Ugot, O.A. (2020). A review of generative adversarial networks and its application in cybersecurity. *Artificial Intelligence Review*, 53, 1721-1736. doi: 10.1007/s10462-019-09717-4.

[49] Zhang, C., Yu, S., Tian, Z., & Yu, J.J. (2023). Generative adversarial networks: A survey on attack and defense perspective. *ACM Computing Surveys*, 56(4), article number 91. doi: 10.1145/3615336.

[50] Zhang, J., Zhong, S., Wang, T., Chao, H.C., & Wang, J. (2020a). Blockchain-based systems and applications: A survey. *Journal of Internet Technology*, 21(1), 1-14. doi: 10.3966/160792642020012101001.

[51] Zhang, X., Ma, Y., Singla, A., & Zhu, X. (2020b). Adaptive reward-poisoning attacks against reinforcement learning. In *Proceedings of the 37th international conference on machine learning* (pp. 11225-11234). San Diego: International Conference on Machine Learning (ICML).

AI vs AI: використання штучного інтелекту для захисту даних та атак на них

Ігор Руденко
Магістр
Національний юридичний університет імені Ярослава Мудрого
61024, вул. Григорія Сковороди, 77, м. Харків, Україна
https://orcid.org/0009-0008-3582-3951

Анотація. Метою дослідження було проведення комплексного аналізу ролі штучного інтелекту як інструменту атак та захисту в інформаційних системах, з акцентом на оцінку ефективності існуючих підходів та обґрунтування перспектив інтеграції штучного інтелекту для посилення кібербезпеки в умовах зростання інтелектуальних загроз. В рамках дослідження було змодельовано протистояння між наступальними та оборонними системами штучного інтелекту в динамічному середовищі з адаптивною поведінкою, що дозволило не лише виявити типові вектори загроз, але й оцінити ефективність відповідних контрзаходів. Було виявлено, що генеративні моделі, зокрема ті, що базуються на навчанні з підкріпленням, ефективно адаптувалися до оборонних реакцій, минаючи традиційні фільтри та евристики. Водночас найвищу стійкість до таких атак продемонстрували комбіновані підходи, що інтегрували федеративне навчання, блокчейн та диференціальну конфіденційність: рівень стійкості до атак збільшився до 40 % при помірному зниженні точності (3-6 %). Змагальне навчання забезпечило підвищення безпеки до 25 %, хоча точність знизилася до 4 %, а його ефективність суттєво залежала від повноти та мінливості навчальних даних. Гомоморфне шифрування виявилося найбільш конфіденційним підходом, але залишалося обмеженим у практичному використанні через надмірне споживання ресурсів та час обробки. Хоча інструменти блокчейну сприяли прозорості та незмінності даних, ці інструменти мали високу затримку, що ускладнювало застосування в умовах реального часу. Загалом, результати дослідження підтвердили доцільність використання мультимодальних, адаптивних та багаторівневих стратегій захисту для систем штучного інтелекту, особливо на тлі зростання кількості генеративних атак, що підтверджується реальними випадками (наприклад, Disney). Практичне значення полягає у формуванні основ для розробки адаптивних систем кіберзахисту, здатних протидіяти інтелектуальним атакам у режимі реального часу. Отримані результати можуть бути використані для підвищення безпеки критичної інфраструктури, фінансових платформ та автономних систем

Ключові слова: конфіденційність даних; генеративні мережі; фішинг; біометрична система; діпфейки