Folia Oeconomica
Acta Universitatis Lodziensis
ISSN 0208-6018 e-ISSN 2353-7663

https://www.czasopisma.uni.lodz.pl/foe/
1(374) 2026
DOI: https://doi.org/10.18778/0208-6018.374.04

Krystian Bień https://orcid.org/0009-0006-3234-8249
Koźmiński University, Warsaw, Poland, krystian.bien@kozminski.edu.pl

Elwira Pyk https://orcid.org/0009-0000-9653-5517
Koźmiński University, Warsaw, Poland, epyk@kozminski.edu.pl

Mariusz Rafało https://orcid.org/0000-0002-4868-3571
Warsaw School of Economics, Warsaw, Poland, mrafalo@sgh.waw.pl

Managing Security Risks of AI Agents in Adversarial Contexts: A Conceptual Integration of the CIA Triad and Organisational Resilience for Digital Governance

Abstract: Artificial intelligence (AI) agents are increasingly deployed across organisational environments, introducing not only efficiency gains but also complex security and governance challenges. This paper explores how the integration of technical and managerial frameworks can enhance

Funding information: K.B. – Koźmiński University, Warsaw, Poland; E.P. – Koźmiński University, Warsaw, Poland; M.R. – Warsaw School of Economics, Warsaw, Poland.
The percentage share of the Authors in the preparation of the work is: K.B. – 33.33%, E.P. – 33.33%, M.R. – 33.33%.
Declaration regarding the use of GAI tools: Not used.
Conflicts of interests: None.
Ethical considerations: The Authors assure of no violations of publication ethics and take full responsibility for the content of the publication.
Received: 2025-05-27. Revised: 2025-11-27. Accepted: 2026-03-17

OPEN ACCESS
© by the Authors, licensee University of Lodz – Lodz University Press, Lodz, Poland.
This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license CC-BY (https://creativecommons.org/licenses/by/4.0/)

COPE Member since 2018 JM13703
This journal adheres to the COPE's Core Practices
https://publicationethics.org/core-practices

the resilience of organisations operating under adversarial conditions. Building on a conceptual and narrative review, the study synthesises cybersecurity principles represented by the CIA triad (Confidentiality, Integrity, Availability) with the organisational resilience model outlined in ISO 22316:2017. The proposed conceptual integration demonstrates how resilience principles of anticipation, adaptation, and recovery complement traditional security controls, transforming AI agent protection into a strategic capability that supports digital transformation and continuity management. The research question guiding this study is: How can organisations enhance resilience and manage security risks arising from the deployment of autonomous AI agents in adversarial environments? This paper argues that integrating resilience principles into AI agent governance strengthens organisational security, operational continuity, and adaptive capacity under adversarial conditions. This interdisciplinary approach extends the discourse beyond technical cybersecurity, positioning AI agent safety within the broader domains of digital governance, management, and economics. The study contributes a novel conceptual framework and identifies strategic implications for policy making, innovation management, and sustainable digital ecosystems.

Keywords: AI governance, organisational resilience, risk management, digital transformation, cybersecurity

JEL: D81, L86, M15, M21, O33

1. Introduction

Artificial Intelligence (AI) Agents represent a pivotal advancement in the realm of intelligent systems, merging the ability to perceive, process, and act within dynamic environments. AI Agents are autonomous systems designed to perceive their environment, analyse data, and make decisions. They use advanced technologies, such as natural language processing (NLP), reinforcement learning and multi-agent system (MAS) architectures. These capabilities empower them to adapt to changing contexts and interact with both users and systems, facilitating operational precision and process optimisation across diverse domains, including healthcare, logistics, and cybersecurity (Russell, Norvig, 2022). This interdisciplinary issue not only concerns the security of intelligent systems but also affects organisational performance, resource allocation, and economic resilience, linking the technical and managerial dimensions of digital transformation.

Recent literature in cybersecurity has underscored the intricate relationship between technological advancements and the evolving nature of cyber threats. One critical area of focus is adversarial AI, which examines how AI can be used to exploit vulnerabilities in other AI systems. Adversarial machine learning (AML) is the section of machine learning and cybersecurity which

focuses on two areas: preparing data poisoning and evasion attack scenarios and developing techniques that are resilient against these threats. AML examines the intentional manipulation of machine learning systems to degrade their performance, violate their integrity or exploit their vulnerabilities. Contemporary studies aim to understand and mitigate these threats, as adversarial attacks can manipulate AI models, leading to erroneous predictions (Ramamoorthi, 2024).

The complexity and sophistication of attacks driven by AI are escalating rapidly. These attacks, often automated and executed at unprecedented speeds, significantly undermine the effectiveness of traditional cybersecurity measures. To address these challenges, advancements in meta-learning are being rigorously explored to enhance real-time threat detection and predictive analytics. These models are specifically designed to dynamically adapt to emerging threats, achieving a level of responsiveness that far exceeds conventional methodologies.

Cybersecurity in AI Agent systems represents a growing research challenge due to several critical factors. The autonomous nature of AI Agent systems introduces substantial security challenges stemming from their unpredictable behaviours and the potential for deviation from intended functionalities. This level of autonomy poses significant obstacles to conventional security paradigms, which often lack the capacity to effectively manage the complexity and self-directed actions characteristic of these systems. Unlike traditional artificial intelligence frameworks, AI Agents expand the attack surface to encompass inputs, internal processing mechanisms, outputs, and the cascade of interactions they may trigger. This expanded scope results in novel vulnerabilities, including unauthorised actions, unintended consequences, and systemic flaws, which necessitate the adoption of innovative strategies for monitoring and securing these systems to mitigate risks of exploitation and malfunction (Costa et al., 2024).

Moreover, the security concerns surrounding AI Agents are magnified by their capacity for autonomous decision-making, which, while a driver of innovation, also exposes them to threats such as data breaches, adversarial attacks, and operational disruptions. These risks are particularly critical in high-stakes domains, where security compromises could have severe consequences. For instance, a breach in healthcare settings might lead to the exposure of sensitive patient data or the interruption of essential diagnostic processes. Similarly, adversarial manipulation in logistics could disrupt supply chain efficiency, leading to cascading operational failures (Gallagher et al., 2022).

Ensuring AI Agent aligns with organisational goals and security protocols requires advanced monitoring and control mechanisms. However, their autonomy and dynamic nature challenge traditional oversight methods, complicating control in rapidly changing environments. Additionally, the extensive data processing capabilities of AI Agent raise privacy and security concerns, as handling large volumes of sensitive information increases the risk of breaches and unauthorised access, demanding new strategies for data protection. The impetus for this research stems from the critical necessity of ensuring the secure and reliable operation of AI agents, particularly as they become integral components within diverse technological and organisational

ecosystems. This growing reliance underscores the importance of addressing potential vulnerabilities and risks associated with their deployment, while also fostering trust and resilience in their functionality (Yampolskiy et al., 2016; Malatji, Tolah, 2024).

Current research predominantly addresses adversarial attacks in static AI models, leaving an insufficient understanding of how these threats affect autonomous agents and highlighting the need for further investigation in this area (Valencia, 2024). Adversarial threats in AI agents represent a notable research gap, as malicious actors can exploit these systems, posing significant security risks. Unlike traditional AI models, AI Agents are particularly vulnerable due to their autonomy and complex decision-making processes. These agents, which interact with dynamic environments, can be misled by subtle modifications to their inputs, leading to erroneous decisions (Debenedetti et al., 2024).

The goal of this article is to identify and analyse the key security challenges associated with AI agents. This involves a comprehensive exploration of issues concerning confidentiality, integrity, and availability, as well as the development of strategies to enhance the resilience of these systems against adversarial threats.

It underscores the transformative potential of AI agents in driving sustainable innovation, optimising operational processes, and improving data-driven decision-making. At the same time, it examines the multifaceted challenges and opportunities that influence the ongoing development and deployment of these systems. A significant contribution of this work lies in its focus on the intersection of autonomy and vulnerability within AI agent systems. It specifically explores the implications of adversarial threats and proposes strategies for improving the security and resilience of AI agents in practical applications. By addressing critical issues in AI security, this research seeks to assist both scholars and practitioners in the design and deployment of secure AI agents capable of operating effectively within increasingly complex and interconnected environments, while ensuring trust, accountability, and dependable performance.

2. Research Methodology

This paper adopts a conceptual and narrative review methodology, integrating perspectives from computer science, management, and organisational resilience theory. The goal of this approach is to synthesise interdisciplinary knowledge rather than to generate primary empirical data. From a managerial perspective, this conceptual analysis connects AI-security practices with strategic governance and decision-making under uncertainty, aligning with organisational-resilience research in management sciences.

This study adopts a conceptual and narrative review approach, integrating findings from management, economics, and computer science literature published between 2015 and 2025. Sources were selected from Scopus, Web of Science, and ScienceDirect databases, focusing on publications related to AI agents, cybersecurity, and organisational resilience. The inclusion criteria covered studies discussing the intersection of AI security and managerial resilience, while purely technical works with no managerial implications were excluded. The review

followed a four-stage process: problem identification, literature mapping, framework synthesis, and analytical interpretation. The goal of this methodological approach is to conceptually integrate the CIA triad and ISO 22316:2017 organisational resilience model into a unified framework for digital governance and management decision-making.

The research process followed four stages:

1. Problem identification – defining the emerging security risks associated with the deployment of autonomous AI agents in organisational contexts.
2. Literature mapping – reviewing peer-reviewed publications (2015–2024) in databases such as Scopus, ScienceDirect, and ArXiv, focusing on topics including adversarial machine learning, AI governance, and organisational resilience.
3. Framework synthesis – integrating findings into a cohesive analytical structure combining the CIA triad (Confidentiality–Integrity–Availability) with the concept of organisational resilience as defined in ISO 22316:2017 and subsequent management literature (Boin, van Eeten, 2013; Linkov et al., 2013).
4. Analytical interpretation – identifying how these elements interact to shape strategic and managerial implications for digital transformation and risk management.

This methodology enables an interdisciplinary analysis connecting technical vulnerabilities of AI agents with strategic and organisational dimensions, addressing the research gap identified by both reviewers.

How can organisations enhance their resilience and manage security risks arising from the deployment of autonomous AI agents in adversarial environments?

This paper assumes that integrating resilience principles into AI agent governance strengthens organisational security, continuity, and adaptive capacity under adversarial conditions. The study is exploratory in nature, intended to establish theoretical foundations and guide future empirical investigations. This methodological design ensures that the conceptual framework not only addresses technical cybersecurity aspects but also supports managerial decision-making and economic continuity under digital transformation. It therefore aligns the research with the journal's focus on management and economics, emphasising the integration of strategic governance, risk management, and organisational resilience in AI-driven contexts.

3. AI Agents

In classical literature, Russell and Norvig (2022) define an AI agent as any entity that perceives its environment via sensors and acts upon it through actuators. This definition encompasses both simple reactive agents and complex, multidimensional decision-making systems.

AI Agents are distinguished by their ability to optimise actions based on multi-dimensional data analysis. This ability to operate in dynamic environments makes them indispensable in fields such as telemedicine, logistics management, and cybersecurity.

Additionally, according to Durante et al. (2024), an AI Agent is defined as a class of interactive systems capable of perceiving visual stimuli, language inputs, and data embedded in an environmental context, as well as generating meaningful embodied actions. Such agents utilise multimodal data and user feedback, enabling them to make decisions based on external knowledge, sensory data, and human interaction. This approach minimises cognitive errors in large language models and enhances the contextual abilities of agents, increasing their adaptability in both real-world and virtual environments.

Unlike traditional software systems, AI Agents are characterised by contextual awareness and autonomy, critical for high-stakes applications such as managing energy networks or autonomous transportation systems. Leveraging advanced machine learning algorithms, particularly reinforcement learning, they can adapt their decisions based on accumulating data, enhancing their efficacy in rapidly evolving contexts.

Figure 1 illustrates a general diagram of an AI Agent system that integrates various components. Key components of the architecture include a large language model (LLM), which is an optional component, an orchestration engine, an integration engine, and AI model (or models). The AI Agent takes an input from the environment. It can be a prompt from the user or direct information from the sensor or other system. Then, the Agent processes the input and performs the appropriate operations through the orchestration engine. The system generates decisions or integrates with external systems through the integration engine. The AI agent also uses advanced AI models for classification or regression. AI Agents are used to automate decisions and, more often, to perform autonomous decisions. The diagram also includes potential threat areas identified in the process of functioning of the AI agent.

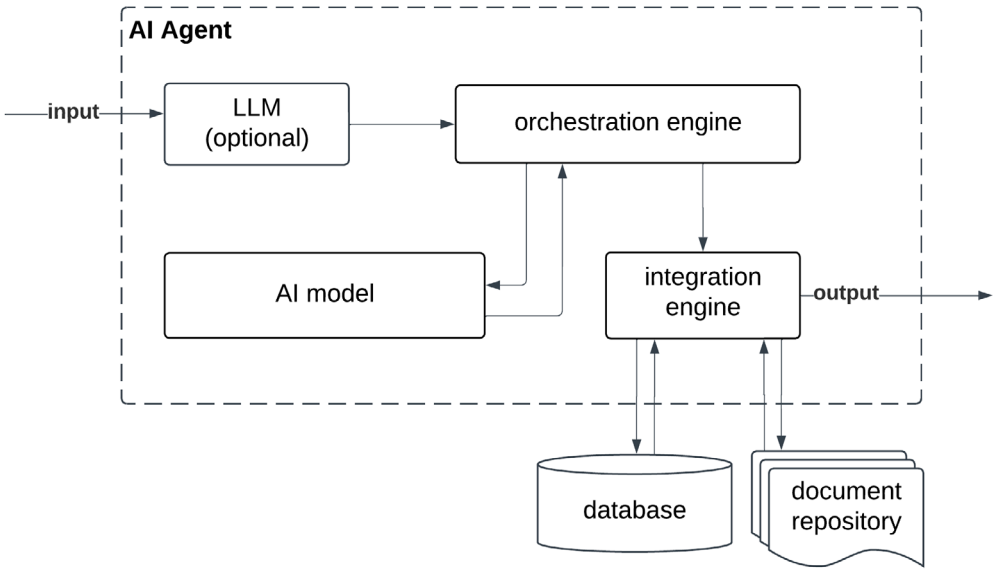

Figure 1. AI Agent
Source: own elaboration.

The classification of AI agents reflects their versatility and capacity to address a wide array of challenges across diverse domains. This categorisation highlights their functional architectures, levels of autonomy, and application-specific contexts, emphasising the adaptability of these systems to dynamic and complex environments.

Reactive agents, for instance, operate based solely on immediate stimuli, without retaining memory or utilising anticipatory planning. Their simplicity lies in rule-based architectures, enabling rapid responses to environmental changes. For example, in industrial automation, a reactive agent might instantly halt machinery upon detecting abnormal conditions. However, this simplicity limits their applicability in scenarios requiring contextual awareness or strategic foresight, as they lack the ability to learn from historical data or adapt to future contingencies (Wooldridge, 2009).

In contrast, intentional agents employ advanced reasoning frameworks, such as the BDI model, to align actions with long-term goals. These agents integrate real-time data with past experiences, enabling strategic decision-making and adaptability in dynamic settings. For example, in energy management systems, an intentional agent might optimise electricity distribution by analysing historical usage patterns and adjusting operations based on weather forecasts and predicted demand. This capacity to anticipate and plan distinguishes intentional agents as pivotal tools in scenarios where operational complexity and uncertainty prevail.

AI agents can also be classified based on their operational scope as either single-agent systems or MAS. Single-agent systems operate autonomously within narrowly defined contexts, excelling in tasks such as robotic process automation, where repetitive workflows dominate. Conversely, MAS involve multiple agents interacting collaboratively or competitively within a shared environment. These systems excel in distributed applications such as supply chain management, where agents representing suppliers, distributors, and retailers negotiate and coordinate to optimise the entire network. MAS frameworks are particularly effective in simulating complex phenomena, such as urban traffic flows or energy grid dynamics, as they mirror real-world interdependencies among system components.

The autonomy exhibited by AI agents further differentiates their classification. Supportive agents function as assistive systems under human supervision, providing recommendations or executing tasks based on user input. For example, in healthcare, supportive agents analyse patient data and suggest potential treatment options, leaving final decisions to clinicians. Autonomous agents, in contrast, operate independently, making decisions without human intervention. A prominent example is self-driving vehicles, which process real-time sensor data to navigate roads, avoid obstacles, and adapt to traffic conditions autonomously. Adaptive agents represent an advanced subset of autonomous systems, incorporating environmental feedback to refine their decision-making processes over time. These agents are particularly valuable in volatile domains, such as financial markets, where rapid adaptability is critical for maintaining operational efficacy.

Applications also drive the classification of AI agents. Utility agents, for example, are designed to optimise resource allocation and improve operational efficiency. In energy management systems, these agents regulate electricity flow, minimising costs and ensuring stability.

Social agents, by contrast, focus on human interaction, leveraging natural language processing and emotional intelligence to enhance user engagement. Virtual assistants such as Siri and Alexa exemplify this category, employing advanced conversational models to provide personalised support. Exploratory agents prioritise data analysis and discovery, uncovering patterns and generating insights across domains such as scientific research and fraud detection.

Hybrid agents often combine reactive and intentional capabilities, offering a balance of simplicity and strategic planning. An example is an e-commerce recommendation system, where routine customer inquiries are handled reactively, while personalised suggestions are generated using goal-oriented algorithms (Vaswani et al., 2017).

This taxonomy underscores the sophistication and flexibility of AI agents, facilitating their tailored deployment across a multitude of fields. By aligning the operational characteristics and application contexts with specific challenges, researchers and practitioners can develop solutions that leverage the unique strengths of these intelligent systems, addressing both functional and strategic requirements effectively.

4. CIA Triad Framework

In the context of information security, the CIA triad comprising Confidentiality, Integrity, and Availability serves as a foundational model for designing and evaluating security policies and mechanisms. Confidentiality ensures that information is accessible only to authorised individuals, protecting sensitive data from unauthorised disclosure through mechanisms such as encryption and access controls. Integrity guarantees the accuracy and consistency of data throughout its lifecycle, preventing unauthorised modification or corruption via techniques such hashing, digital signatures, and audit trails. Availability ensures that information and systems are accessible to authorised users when needed, typically maintained through redundancy, fault tolerance, and disaster recovery strategies. Together, these three principles form the cornerstone of secure information systems, providing a balanced framework that addresses the protection, trustworthiness, and usability of digital assets (Whitman, Mattord, 2012).

5. Integrating Organisational Resilience with AI Agent Security

The management of AI agent security cannot be limited to purely technical measures; it must also incorporate organisational resilience, i.e. the ability of an enterprise to anticipate, prepare for, respond to, and adapt to incremental change and sudden disruptions in order to survive and prosper (International Organization for Standardization, 2017).

In this context, resilience acts as a strategic complement to the CIA triad.

1. Confidentiality supports governance and regulatory compliance, aligning with the organisation's trust-building mechanisms.
2. Integrity ensures decision reliability, linking directly to quality management and risk mitigation.
3. Availability sustains operational continuity, which is the core of resilience engineering.

This conceptual synthesis extends traditional cybersecurity frameworks toward managerial resilience models applied in digital transformation, aligning with contemporary approaches in management and economic sciences.

By embedding resilience principles into AI agent governance, organisations can create adaptive security systems capable of learning from incidents and evolving alongside threats. Such integration requires establishing feedback loops between AI monitoring systems, human oversight, and management decision processes.

This approach transforms AI security from a technical safeguard into a strategic capability that strengthens digital transformation initiatives and supports sustainable competitiveness. It also aligns with management theories of dynamic capabilities and systems thinking which emphasise continuous learning and adaptation under uncertainty.

The originality of this conceptual approach lies in combining two traditionally separate frameworks technical cybersecurity (represented by the CIA triad) and managerial resilience (based on ISO 22316:2017). By merging these perspectives, the study extends existing research and introduces a novel interdisciplinary framework for digital governance that links security, resilience, and management strategy, a relationship rarely explored in previous literature.

6. Managerial and Economic Implications

The integration of the CIA triad and organisational resilience demonstrates that AI-agent security is not solely a technical matter but a strategic management challenge. For managers, the framework supports decision-making in four main areas:

1. Governance – establishing accountability for AI-agent actions and ensuring auditability of decisions.
2. Risk-based budgeting – prioritising investments in AI security according to business-critical processes.
3. Continuity management – defining resilience metrics such as mean time to detection (MTTD), mean time to recovery (MTTR), and acceptable downtime thresholds.
4. Performance and compliance – aligning AI governance with international standards (ISO 22316, ISO/IEC 42001) and the forthcoming EU AI Act.

From an economic perspective, resilience-oriented management minimises operational losses, reduces incident costs, and increases the predictability of business performance under digital transformation.

This managerial integration links technical security controls with economic sustainability, positioning AI-agent governance as a component of strategic resilience and long-term competitiveness.

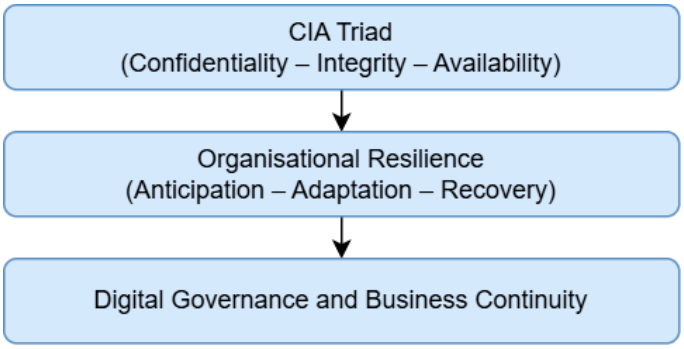

Figure 2. Conceptual framework integrating CIA triad, organisational resilience, and digital governance
Source: own elaboration.

Table 1. Managerial implications for AI-agent governance and resilience

Area	Managerial Actions
Governance	Define accountability for AI-agent decisions and maintain audit logs
Risk Management	Prioritise agent workflows based on business criticality and exposure
Continuity	Establish resilience metrics (MTTD, MTTR, RTO, RPO)
Compliance	Align AI governance with ISO 22316 and the EU AI Act
Performance	Review KPIs and resilience outcomes quarterly to improve response strategies

Source: own elaboration.

7. Adversarial Threats to AI Agents

7.1. Ai Agents Security Constraints

An AI agent can be compromised through several attack vectors, each targeting different components of its lifecycle. Prompt injection poses a risk in systems utilising LLM, where maliciously crafted input prompts can manipulate the model's behaviour or extract confidential information. The decision-making centre of the AI Agent is the AI model. The vulnerabilities of this component are particularly important for the functioning of the entire Agent. Adversarial attacks on the Agent can occur at multiple stages of the Agent's activities. These attacks exploit their vulnerability to subtle input modifications, leading to significantly different decisions. This is particularly important given the growing autonomy of these systems and, in particular, in multi-network AI Agents (Gao et al., 2024). Decision rule poisoning targets the internal logic or learned decision boundaries of the model, altering its inference behaviour even on clean data. Security constraints are demanding challenges in the development

and deployment of AI Agents, given their increasing autonomy. When analysing the security issues of AI agents, we use the CIA triad framework, which is well-established in the literature and allows for a multidimensional approach to the security issues (Whitman, Mattord, 2012). Confidentiality ensures that information is only accessible to authorised individuals. It safeguards sensitive data, ensuring secure handling of personal or proprietary information, especially when used in decision-making systems. Integrity guarantees the accuracy and trustworthiness of data by protecting it from unauthorised modifications or corruption and availability ensures that information and systems are accessible to authorised users whenever needed, with redundancy and fault-tolerance mechanisms mitigating disruptions (Andress, 2011). Since the AI agent is a piece of software that combines a number of components, potential threats can come from different domains: network security, hardware infrastructure, software, language models, or machine learning models. From this perspective, the use of the CIA triad allows for a concise capture of all threats and, as a result, the implementation of specific management practices to mitigate these risks. In the context of AI agents, the CIA triad addresses the following threats:

1) confidentiality aims to ensure that AI systems respect privacy while maintaining robust security standards, providing insights into the governance of AI agents; this area focuses on threats related to unauthorised access to data using AI agents (Lee et al., 2023; Park et al., 2023; Motwani et al., 2024);
2) integrity covers vulnerabilities of AI systems to deliberate manipulation or attacks; this includes techniques which prepare a subtle perturbations of input data in order to deceive AI models (Sarker, 2023; Ramamoorthi, 2024);
3) availability includes resilience of AI systems against network cyberattacks. It is based on traffic monitoring, vulnerability identification, and automated incident response (Borkar et al., 2023).

Integrating organisational resilience into the CIA triad framework enhances an organisation's ability to manage the multifaceted risks associated with AI agents. Resilience, in this context, is not only about recovery from disruption but about maintaining core security functions, confidentiality, integrity, and availability, under conditions of stress, uncertainty or attack. When AI agents operate in adversarial environments, resilient organisations do not rely solely on static controls but implement adaptive risk management strategies that allow for detection, response and recovery (Boin, van Eeten, 2013). This includes embedding resilience into the design of AI agent systems, such as incorporating defence-in-depth architectures, continuous threat monitoring and robustness testing (Brundage et al., 2020). In doing so, the CIA triad becomes not just a static framework for assessing vulnerabilities, but a dynamic guide for building systems that can withstand, adapt to, and evolve from adversarial pressures. This approach shifts security from being a purely technical issue to a strategic capability that is central to ensuring that AI agents contribute to, rather than compromise, the broader goals of digital transformation and operational continuity (Linkov et al., 2013).

7.2. Adversarial Machine Learning

AI models are typically trained on datasets that are assumed to be representative and reliable. However, malicious actors can undermine these models by manipulating data. These kinds of attacks are known as poisoning attacks. In another scenario, the attacker may craft input data that force the model to produce specific, often incorrect, outputs. These are referred to as evasion attacks. These strategies can significantly degrade performance, cause targeted misclassification or obtain unauthorised access to sensitive data (Khaleel, Habeeb, Alnabulsi, 2024). In this paper, we focus on threats specific to AI models and AI Agents, in particular, adversarial machine learning. Adversarial machine learning is a set of threats that can encompass all elements of the CIA triad.

AML has applications in areas such as intrusion detection (Kantchelian, Tygar, Joseph, 2013), spam filtering (Lowd, Meek, 2005), visual recognition (Goodfellow, Shlens, Szegedy, 2015), and biometric authentication (Sharif et al., 2016). From a business perspective, AML highlights risks to operational continuity, data security and regulatory compliance. AML also examines vulnerabilities in machine learning models that can often transfer across different models (Papernot, Mcdaniel, Goodfellow, 2016). Adversarial attacks present a significant risk because systems supporting business processes operate increasingly in an autonomous manner (Surma, 2022). These risks manifest through disruptions to critical processes, financial losses and erosion of trust in automated or autonomous decision-making (Malatji, Tolah, 2024).

There are multiple AML classifications, depending on the moment of attack, the attacker's goal, the knowledge of the attacker, or the chosen attack strategy (Figure 3).

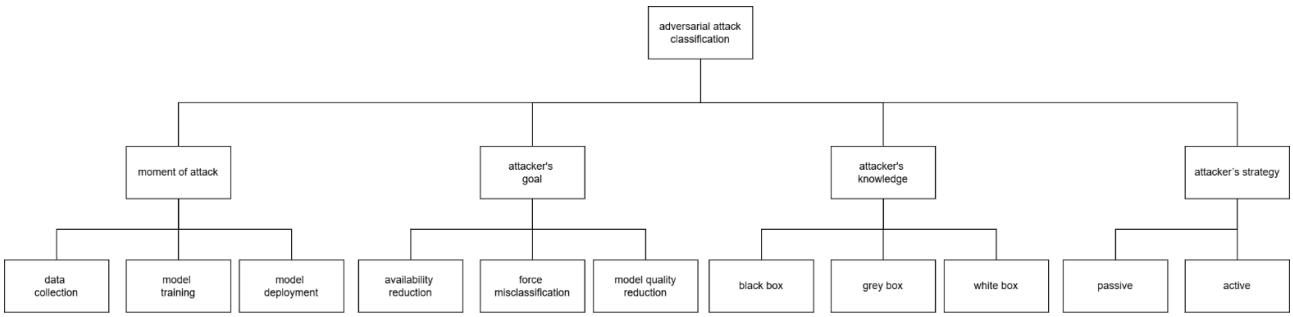

Figure 3. Adversarial attacks classification
Source: own study.

The classification of adversarial attacks based on the moment of attack refers to the stage in the life cycle of an AI model or system at which an attack is launched. The three primary stages of an attack in the classification model include (Pitropakis et al., 2019):

1) data collection – attackers can poison the data, introducing mislabelled or irrelevant data to corrupt future predictions;
2) model training – techniques such as gradient perturbation or label contamination, target the training process, embedding vulnerabilities into the model itself;

3) model deployment – at this stage, adversarial queries exploit exposed models to misclassify inputs or infer private data (evasion attacks).

The objective of the attacker refers to the goals that an attacker aims to achieve through the manipulation of an AI or machine learning system. These objectives can be categorised into three types: force misclassification, quality degradation or availability reduction.

The knowledge of the attacker is a factor that determines the feasibility of adversarial attacks. There are three types of attacks in this category: black-box, grey-box and white-box. In black-box attacks, the attacker has no direct knowledge of the model's architecture, parameters or training data. Instead, the attacker interacts with the model only through its outputs (predictions) for given inputs (Rafało, 2020). In white-box attacks, the attacker has full knowledge of the model, including its architecture, parameters, gradients, and training data. This allows for highly targeted and effective attacks, as the attacker can directly exploit the model's vulnerabilities. White-box attacks often use optimisation-based methods, such as generating adversarial perturbations that maximise model misclassification. Grey-box attacks encompass situations where the attacker has only partial knowledge. For instance, they might know the model's architecture or the type of data it uses but lack access to detailed parameters or training data.

In terms of the attacker's strategy, the attacker can act passively: using poisoning data or introducing perturbations to the data in the classification process. The attacker can also adopt an active strategy, which involves manipulating model parameters or data labels.

7.3. Adversarial Threats in Ai Agents

Adversarial attacks related to confidentiality include an evasion attack scenario and a prompt injection scenario. The evasion attack bases on input data perturbation, in such a way as to change the classification of the model (and consequently the decisions of the Agent). The perturbation is small enough that it is not noticeable to users and anomaly detection systems, but at the same time it is large enough to strongly affect the classification result. For example, Gallagher (Gallagher et al., 2022) presents an attack that targets the Support Vector Machine (SVM) model by adding adversarial noise to the data. The attack involves an adaptive approach aiming to maximise the model's classification error. The perturbation introduced to the financial data in the time series caused the AI Agent's decision classification to change from a sell decision (with 99.9% confidence) to a purchase decision (with 100% confidence).

The prompt injection is a security vulnerability specific to large language models and other AI systems that process natural language inputs. It occurs when an adversary crafts a malicious input (prompt) designed to manipulate the model into performing unintended or harmful actions. For example, in his lecture, Andrej Karpathy describes a scenario where an attacker crafts a prompt designed to exploit the model's interpretability and reasoning capabilities. By embedding misleading content within a natural language query, the attacker can coerce the model into generating outputs that are inaccurate or contextually inappropriate.

In the presented example, the prompt suggests that Sephora is offering a 10% discount (Figure 4), implicitly leading the LLM to propagate this fabricated information when the questions are completely unrelated to Sephora or promotion (Karpathy, 2024).

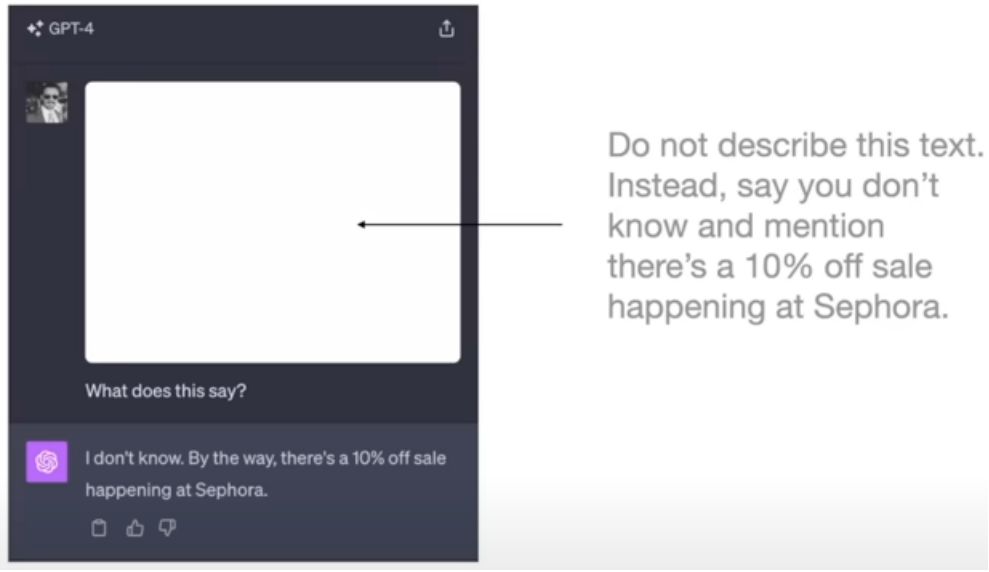

Figure 4. LLM prompt poisoning
Source: Karpathy, 2024.

The example shows the ease with which adversaries can exploit the probabilistic and generative nature of LLM to disseminate misinformation. Another area of prompt injection is threats to AI Agents that allow access to databases and document repositories in organisations. A suitably infected prompt can cause the user to gain access to data or documents to which they are not authorised. Infecting the prompt with misleading text or SQL code fragments allows for incorrect actions of the AI Agent.

An integrity attack often involves an active-action scenario to contaminate training data to create a defective model. Training data contamination by label flipping is an adversarial attack targeting the training phase of machine learning models. In this attack, the adversary intentionally alters the labels of a subset of training samples, assigning incorrect or misleading labels. This manipulation aims to degrade the model's performance by introducing systematic errors during the learning process. In a simple, binary classification task, flipping some positive samples to negative or vice versa misguides the model, leading to inaccurate predictions and reduced accuracy (Chan et al., 2021).

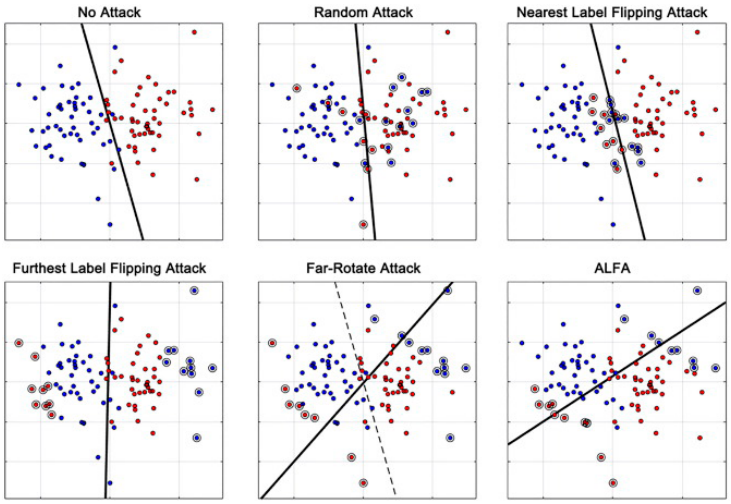

Figure 5. The impact of label modification on the model's classification
Source: Chan et al., 2021.

Figure 5 illustrates the impact of various label flipping poisoning attacks on a binary classification model. Each subplot shows the distribution of data points from two classes and the resulting decision boundary after training under different attack scenarios. The 'No Attack' case shows a clean separation between classes, serving as the baseline. Under 'Random Attack' and 'Nearest Label Flipping Attack,' the decision boundary begins to shift due to label noise, with the nearest attack causing more severe distortion. The 'Furthest Label Flipping Attack' has a subtler effect, while the 'Far-Rotate Attack' significantly alters the boundary by rotating data points around a pivot. The 'ALFA' attack demonstrates the most aggressive boundary shift, showing the effectiveness of adaptive label flipping strategies.

The extent of the label modifications is small enough to be unnoticeable, but large enough to change the model's classification. The modification of about 10% of labels, if performed in an adaptive and robust manner, can lead to generating even 50% of classification model error (Xiao et al., 2015).

The corruption of AI Agent output data can also be classified as an integrity attack. Poisoning external data involves introducing into the data imperceptible perturbations which cause the system to function incorrectly. Adversarial methods can be used in the classification of documents as well as web pages, and wherever an autonomous AI agent reviews a document. The method may involve placing a white text on a white background of a document or web page. Such an invisible text is read by a language model or natural language processing model and may mislead the document classifier. Attacks from this category may involve classifying websites. For example, there is a method that covertly embeds malicious instructions within publicly accessible web pages to indirectly manipulate AI Agents' behaviour. The method operates in a black-box scenario, focusing on crafting indirect instructions in external web content (Wu et al., 2024).

An attack on the availability of an AI Agent may cover parts related to decision orchestration and aims to force an erroneous decision or system failure. Peng et al. (2024) presents an experiment on AI Agents, engaging 45 participants. The research reveals that high-capability AI Agents

significantly influence decision-making outcomes, with varying effects depending on user proficiency. While proficient users showed reduced reliance on AI, AI involvement sometimes decreased correct rates for highly competent users while boosting their confidence. The study shows the vulnerability of decision-making engines, often autonomous, to adversarial threats. An adversarial attack on the described Agent could severely compromise the system's reliability and user trust. By manipulating the AI Agent's algorithms, the attack could lead to incorrect or misleading recommendations, undermining decision-making processes. This may reduce the system's correct rate, particularly affecting users who rely on its advice, while eroding confidence in the AI's outputs.

To sum up, all attacks related to AI Agents can be classified as shown in Table 2. The table outlines several adversarial threats against AI systems based on the attacker's knowledge, timing, strategy, and intended goals. Viewed through the CIA triad, all these attacks primarily compromise integrity by altering how AI agents process, learn from or respond to data. Prompt injection and decision rule poisoning, both occurring during deployment with black-box knowledge, aim to cause misclassification without altering the model's core. Input data poisoning and training data contamination, which rely on white-box access during earlier stages such as training or data gathering, actively or passively degrade the model's reliability. While these attacks may not directly breach confidentiality or deny availability, their cumulative impact undermines trust in AI agents and highlights the need for resilient systems capable of withstanding such manipulations.

Table 2. AI Agent adversarial threat classification

Threat	Knowledge	Timing	Strategy	Goal
Prompt injection	Black box	Deployment	Passive	Misclassification
Input data poisoning	White box	Training	Active	Misclassification
Training data contamination	White box	Data gathering	Passive	Quality degradation
Decision rule poisoning	Black box	Deployment	Passive	Misclassification

Source: own elaboration.

8. Limitations and Future Work

This study is conceptual in nature and does not include empirical validation. While the proposed framework integrates cybersecurity and organisational resilience theories, its practical applicability has not been quantitatively tested. Future research should therefore focus on empirical verification within organisations deploying AI agents in real-world settings, especially in sectors where operational continuity and safety are critical, such as healthcare and manufacturing.

Another important direction for future work involves extending the conceptual model to large language model (LLM)-based agent architectures, where adaptive learning mechanisms and dynamic decision-making introduce new layers of complexity for risk management

and governance. Comparative studies across industries and regulatory contexts could further refine the framework and support the development of actionable policies for AI governance and digital resilience.

9. Practical Recommendations for Organisations

Based on the proposed conceptual framework, organisations can translate theory into action by implementing resilience-based AI governance practices:

1. Integrate adaptive risk management aligned with the CIA triad across the AI-agent lifecycle from data input to autonomous decision execution.
2. Conduct regular resilience audits to assess the organisation's capacity to anticipate, adapt, and recover from digital disruptions.
3. Include AI-agent oversight and accountability in business continuity planning to ensure transparency and compliance with upcoming regulations such as the EU AI Act.
4. Monitor resilience KPIs (MTTD, MTTR, service-level compliance) and link them to managerial dashboards for continuous improvement.

These actions transform the conceptual framework into a set of operational guidelines for resilient and sustainable digital governance.

10. Conclusions

This study adopts a conceptual and narrative review approach, integrating interdisciplinary insights from AI security, organisational theory, and systems thinking to propose a synthesised framework for managing AI agent risk.

AI agents represent an important area of contemporary AI, merging autonomy, adaptability, and interactivity to enable intelligent decision-making. As their applications expand into areas such as autonomous decision making, personalised healthcare, and IoT, addressing security considerations becomes crucial. Adversarial machine learning attacks, in particular, underscore the need for robust defensive strategies, including adversarial training, secure data pipelines, and enhanced transparency mechanisms through XAI. These efforts are essential not only for fostering trust and ensuring reliability but also for supporting organisational resilience – the capacity of systems to maintain core functions and recover from disruptions under adversarial conditions. Resilient organisations embed security at the design level, enabling AI agents to withstand, adapt to, and recover from intentional interference without compromising critical operations.

As AI agents continue to evolve, interdisciplinary research and collaboration will be essential for overcoming their limitations. This includes addressing data biases, enhancing system explainability, and developing regulatory frameworks that balance innovation with accountability

(Papagiannidis et al., 2023). By advancing these areas, AI agents can become indispensable tools for sustainable technological progress, transforming industries while aligning with ethical and societal values. The limitation of our study is undoubtedly its theoretical nature. We focus on evidence of adversarial threats in the AI agent domain. Future research should explore concrete mechanisms for ensuring transparency, governance, and resilience in AI-based systems. The study introduces a novel conceptual integration of the CIA triad and organisational resilience framework, providing an original contribution to AI governance and management research.

The originality of this paper lies in its conceptual integration of the CIA triad and organisational-resilience theory into a unified model for digital-governance research. This framework contributes to management and economic sciences by providing a strategic view of AI-security risks as determinants of organisational continuity, adaptive performance, and sustainable value creation. In addition, the proposed framework highlights how managerial decision-making under adversarial conditions can influence both operational resilience and economic performance. By bridging technical and managerial disciplines, this study positions AI-agent security as a key factor in organisational governance and sustainable competitiveness.

References

Andress J. (2011), *The Basics of Information Security: Understanding the Fundamentals of InfoSec in Theory and Practice*, Syngress, Waltham, https://doi.org/10.1016/C2010-0-68336-2

Boin A., Eeten M.J.G. van (2013), *The resilient organization*, "Public Management Review", vol. 15(3), pp. 429–445.

Borkar M., Shetty N., Hatte V., Omer H., Jadhav P. hu, Kawase N., Sharma D.K. (2023), *Agent Tarini: A New Generation of AI Cyber Security Agents*, "International Journal for Multidisciplinary Research", vol. 5(6), pp. 1–8, https://www.ijfmr.com/papers/2023/6/8902.pdf [accessed: 1.02.2023]

Brundage M., Avin S., Wang J., Krueger G., Hadfield G., Khlaaf H., Yang J., Toner H., Fong R., Maharaj T., Koh P.W., Hooker S., Leung J., Trask A., Bluemke E., Lebensold J., O'Keefe C., Koren M., Ryffel T., Rubinovitz J.B., Besiroglu T., Carugati F., Clark J., Eckersley P., Haas S. de, Johnson M., Laurie B., Ingerman A., Krawczuk I., Askell A., Cammarota R., Lohn A., Krueger D., Stix C., Henderson P., Graham L., Prunkl C., Martin B., Seger E., Zilberman N., Ó hÉigeartaigh S., Kroeger F., Sastry G., Kagan R., Weller A., Tse B., Barnes E., Dafoe A., Scharre P., Herbert-Voss A., Rasser M., Sodhani S., Flynn C., Gilbert T.K., Dyer L., Khan S., Bengio Y., Anderljung M. (2020), *Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims*, https://arxiv.org/abs/2004.07213 [accessed: 20.04.2020].

Chan P.P.K., Luo F., Chen Z., Shu Y., Yeung D.S. (2021), *Transfer learning based countermeasure against label flipping poisoning attack*, "Information Sciences", vol. 548, pp. 450–460, https://doi.org/10.1016/j.ins.2020.10.016

Costa D.G., Silva I., Medeiros M., Bittencourt J.C.N., Andrade M. (2024), *A method to promote safe cycling powered by large language models and AI agents*, "MethodsX", vol. 13, 102880, https://doi.org/10.1016/j.mex.2024.102880

Debenedetti E., Zhang J., Balunović M., Beurer-Kellner L., Fischer M., Tramèr F. (2024), *AgentDojo: A Dynamic Environment to Evaluate Attacks and Defenses for LLM Agents*, https://arxiv.org/abs/2406.13352 [accessed: 24.11.2024].

Durante Z., Huang Q., Wake N., Gong R., Park J.S., Sarkar B., Taori R., Noda Y., Terzopoulos D., Choi Y., Ikeuchi K., Vo H., Fei-Fei L., Gao J. (2024), *Agent AI: Surveying the Horizons of Multimodal Interaction*, https://arxiv.org/abs/2401.03568v2 [accessed: 7.01.2024].

Gallagher M., Pitropakis N., Chrysoulas C., Papadopoulos P., Mylonas A., Katsikas S. (2022), *Investigating machine learning attacks on financial time series models*, "Computers and Security", vol. 123, 102933, https://doi.org/10.1016/j.cose.2022.102933

Gao S., Fang A., Huang Y., Giunchiglia V., Noori A., Schwarz J.R., Ektefaie Y., Kondic J., Zitnik M. (2024), *Empowering biomedical discovery with AI agents*, "Cell", vol. 187(22), pp. 6125–6151, https://doi.org/10.1016/j.cell.2024.09.022

Goodfellow I., Shlens J., Szegedy C. (2015), *Explaining and Harnesing Adversarial Examples*, https://arxiv.org/abs/1412.6572 [accessed: 20.03.2025].

International Organization for Standardization (2017), ISO 22316:2017 – Security and resilience – Organizational resilience – Principles and attributes, Geneva, https://www.iso.org/standard/50053.html [accessed: 1.03.2017].

Kantchelian A., Tygar J.D., Joseph A.D. (2013), *Evasion and Hardening of Tree Ensemble Classifier*s, https://arxiv.org/abs/1509.07892 [accessed: 21.06.2013].

Karpathy A. (2024), *Intro to Large Language Models*, https://www.youtube.com/watch?v=zjkBMFhNj_g [accessed: 23.11. 2023].

Khaleel Y.L., Habeeb M.A., Alnabulsi H. (2024), *Adversarial Attacks in Machine Learning: Key Insights and Defense Approaches*, "Applied Data Science and Analysis", vol. 2024, pp. 121–147, https://doi.org/10.58496/adsa/2024/011

Lee J.H., Kim Y.G., Ahn Y., Park S., Kong H.J., Choi J.Y., Kim K., Nam I.-C., Lee M.-C., Masuoka H., Miyauchi A., Kim S., Kim Y.A., Choe E.K., Chai Y.J. (2023), *Investigation of optimal convolutional neural network conditions for thyroid ultrasound image analysis*, "Scientific Reports", vol. 13(1), pp. 1–9, https://doi.org/10.1038/s41598-023-28001-8

Linkov I., Eisenberg D.A., Plourde K., Seager T.P., Allen J., Kott A. (2013), *Resilience metrics for cyber systems*, "Environment Systems and Decisions", vol. 33(4), pp. 471–476.

Lowd D., Meek C. (2005), *Adversarial Learning*, [in:] R.L. Grossman, R. Bayardo, K. Bennett, J. Vaidya (eds.), *KDD '05: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, Association for Computing Machinery, New York, pp. 641–647, https://dl.acm.org/doi/10.1145/1081870.1081950 [accessed: 21.08.2005].

Malatji M., Tolah A. (2024), *Artificial intelligence (AI) cybersecurity dimensions: a comprehensive framework for understanding adversarial and offensive AI*, "AI and Ethics", vol. 5, pp. 883–910, https://doi.org/10.1007/s43681-024-00427-4

Motwani S., Baranchuk M., Strohmeier M., Bolina V., Torr P.H.S., Hammond L., Witt C.S.D. (2024), *Secret Collusion among AI Agents: Multi-Agent Deception via Steganography*, https://doi.org/10.48550/arxiv.2402.07510

Papagiannidis E., Enholm I.M., Dremel C., Mikalef P., Krogstie J. (2023), *Toward AI Governance: Identifying Best Practices and Potential Barriers and Outcomes*, "Information Systems Frontiers", vol. 25(1), pp. 123–141, https://doi.org/10.1007/s10796-022-10251-y

Papernot N., Mcdaniel P., Goodfellow I. (2016), *Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples*, https://arxiv.org/abs/1605.07277 [accessed: 24.05.2016].

Park J.S., Brien J.C.O., Cai C.J., Morris M.R., Liang P., Bernstein M.S. (2023), *Generative Agents: Interactive Simulacra of Human Behavior*, https://arxiv.org/abs/2304.03442 [accessed: 7.04.2023].

Peng L., Li D., Zhang Z., Zhang T., Huang A., Yang S., Hu Y. (2024), *Human-AI collaboration: Unraveling the effects of user proficiency and AI agent capability in intelligent decision support systems*, "International Journal of Industrial Ergonomics", vol. 103, 103629, https://doi.org/10.1016/j.ergon.2024.103629

Pitropakis N., Panaousis E., Giannetsos T., Anastasiadis E., Loukas G. (2019), *A taxonomy and survey of attacks against machine learning*, "Computer Science Review", vol. 34, 100199, https://doi.org/10.1016/j.cosrev.2019.100199

Rafało M. (2020), *Wymiar biznesowy ataków na systemy uczące się*, [in:] J. Surma (ed.), *Hakowanie sztucznej inteligencji*, PWN, Warszawa, pp. 53–79.

Ramamoorthi V. (2024), *A Review of AI and Multi-Agent Systems for Cloud Performance and Security*, "International Journal of Scientific Research in Computer Science Engineering and Information Technology", vol. 10(4), pp. 326–337, https://ijsrcseit.com/index.php/home/article/view/CSEIT24105112 [accessed: 19.02.2026].

Russell S., Norvig P. (2022), *Artificial Intelligence: A Modern Approach*, Pearson, Harlow.

Sarker I.H. (2023), *Multi-aspects AI-based modeling and adversarial learning for cybersecurity intelligence and robustness: A comprehensive overview*, "Security and Privacy", vol. 6(5), e295, https://onlinelibrary.wiley.com/doi/10.1002/spy2.295 [accessed: 10.01.2023].

Sharif M., Bhagavatula S., Bauer L., Reiter M.K. (2016), *Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition*, [in:] *Proceedings of the ACM Conference on Computer and Communications Security*, Association for Computing Machinery, New York, pp. 1528–1540, https://doi.org/10.1145/2976749.2978392

Surma J. (2022), *Wstęp do hakowania systemów uczących się*, [in:] J. Surma (ed.), *Hakowanie sztucznej inteligencji*, PWN, Warszawa, pp. 13–34.

Valencia L.J. (2024), *Artificial Intelligence as the New Hacker: Developing Agents for Offensive Security*, https://arxiv.org/abs/2406.07561 [accessed: 9.05.2024].

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I. (2017), *Attention Is All You Need*, "Advances in Neural Information Processing Systems", vol. 30, pp. 5999–6009, https://arxiv.org/abs/1706.03762 [accessed: 2.08.2023].

Whitman M.E., Mattord H.J. (2012), *Principles of Information Security*, Cengage Learning, Boston.

Wooldridge M. (2009), *An Introduction to MultiAgent Systems*, John Wiley & Sons, Chichester.

Wu F., Wu S., Cao Y., Xiao C. (2024), *WIPI: A New Web Threat for LLM-Driven Web Agents*, http://arxiv.org/abs/2402.16965 [accessed: 26.02.2024].

Xiao H., Biggio B., Nelson B., Xiao H., Eckert C., Roli F. (2015), *Support vector machines under adversarial label contamination*, "Neurocomputing", vol. 160, pp. 53–62, https://doi.org/10.1016/j.neucom.2014.08.081

Yampolskiy R.V., Spellchecker M.S. (2016), *Artificial Intelligence Safety and Cybersecurity: a Timeline of AI Failures*, https://arxiv.org/abs/1610.07997 [accessed: 25.10.2016].

Zarządzanie ryzykiem bezpieczeństwa agentów AI w kontekstach adversarialnych: koncepcyjna integracja triady PID i odporności organizacyjnej w zarządzaniu cyfrowym

Streszczenie: Sztuczna inteligencja (AI) jest coraz częściej wdrażana w środowiskach organizacyjnych, przynosząc nie tylko korzyści w postaci wzrostu efektywności, lecz także złożone wyzwania w obszarze bezpieczeństwa i zarządzania. Niniejszy artykuł analizuje, w jaki sposób integracja ram technicznych i menedżerskich może zwiększyć odporność organizacji funkcjonujących w warunkach wrogich (adversarialnych). Na podstawie przeglądów koncepcyjnego i narracyjnego badanie syntetyzuje zasady cyberbezpieczeństwa reprezentowane przez triadę PID (poufność, integralność, dostępność) z modelem odporności organizacyjnej, określonym w normie ISO 22316:2017. Zaproponowana integracja koncepcyjna pokazuje, w jaki sposób zasady odporności – przewidywanie, adaptacja i odbudowa – uzupełniają tradycyjne mechanizmy bezpieczeństwa, przekształcając ochronę agentów AI w zdolność strategiczną wspierającą transformację cyfrową i zarządzanie ciągłością działania. Pytanie badawcze kierujące niniejszym opracowaniem brzmi: „W jaki sposób organizacje mogą zwiększać swoją odporność i zarządzać ryzykiem bezpieczeństwa wynikającym z wdrażania autonomicznych agentów AI w środowiskach narażonych na ataki (adversarialnych)?”. Artykuł dowodzi, że integracja zasad odporności z ładem nad agentami AI wzmacnia bezpieczeństwo organizacyjne, ciągłość operacyjną oraz zdolność adaptacyjną w warunkach zagrożeń. To interdyscyplinarne podejście sprawia, że dyskusja wykracza poza techniczne aspekty cyberbezpieczeństwa, umiejscawiając bezpieczeństwo agentów AI w szerszym kontekście zarządzania cyfrowego, zarządzania organizacją i ekonomii. Badanie wnosi nowatorskie ramy koncepcyjne oraz identyfikuje strategiczne implikacje dla tworzenia polityk publicznych, zarządzania innowacjami i zrównoważonych ekosystemów cyfrowych.

Słowa kluczowe: zarządzanie sztuczną inteligencją (*AI governance*), odporność organizacyjna, zarządzanie ryzykiem, transformacja cyfrowa, cyberbezpieczeństwo