The CSI Journal on Computer Science and Engineering

Journal homepage: www.csionjcse.ir

A Quantitative Adversarial Framework for Cybersecurity Risk Analysis of AI-Enabled Systems

Mehdi Fazli[1*]

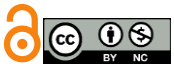
[1] Department of Mathematics, Ardabil Branch, Islamic Azad University, Ardabil, Iran

*** Corresponding author email address**: mehdi.fazli.s@gmail.com

A r t i c l e I n f o

Article type:
Original Research

How to cite this article:
Fazli, M. (2026). A Quantitative Adversarial Framework for Cybersecurity Risk Analysis of AI-Enabled Systems. *The CSI Journal on Computer Science and Engineering, 20*(2), 136-151.
http://dx.doi.org/10.22034/jcse.2026.590576.1088



Copyright: © 2026 by the authors. Published under the terms and conditions of Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License.

A B S T R A C T

The increasing integration of artificial intelligence into safety-critical and large-scale digital systems challenges conventional cybersecurity risk analysis methods. Traditional approaches struggle to capture AI-specific characteristics such as emergent behavior, vulnerability to adversarial machine learning attacks, and the growing overlap between security, safety, and accountability concerns. At the same time, regulatory initiatives highlight the need for structured, quantitative risk analysis methods for high-risk AI systems. This paper proposes an enhanced cybersecurity risk analysis framework tailored to systems that incorporate AI components. Building on adversarial risk analysis, the framework explicitly models AI-related impacts, learning-based assets, intelligent security and recovery controls, and AI-enabled targeted attacks. System architectures are decomposed into hierarchical blocks, enabling probabilistic simulation of attack entry, propagation, defensive failure, and both local and systemic impacts. Strategic attacker behavior is captured by modeling adversarial decision-making under uncertainty and integrating it into Monte Carlo based risk estimation. The applicability of the framework is demonstrated through a case study involving automated driving systems. The results show that incorporating AI-specific defenses and strategic attacker modeling leads to substantial reductions in tail risk and can induce measurable deterrence effects. The proposed approach supports risk-aware cybersecurity investment decisions and provides an analytical foundation for managing and accessing high-risk AI systems in regulated environments.

Keywords: *High-Risk AI Systems; Adversarial Risk Analysis; Artificial Intelligence; Adversarial Machine Learning*

Article history:
Published online 07 August 2026

1 Introduction

Artificial intelligence (AI) has rapidly become a central driver of technological transformation across a wide range of domains. Its ability to accelerate processes that previously required years of costly research such as drug discovery highlights its disruptive potential. At the same time, AI has enabled the development of radically new technologies, including automated driving systems and unmanned aerial vehicles, while also finding widespread adoption in major sectors such as education, banking, and financial services. These advances clearly demonstrate the strategic importance of AI for modern economies and societies[1-5].

Nevertheless, alongside these benefits, AI also introduces substantial risks. Notable examples include the potential misuse of AI systems for the generation of biochemical weapons or the exploitation of deepfake technologies to manipulate identities, violate privacy, and cause reputational harm. These concerns are further amplified by the rapid pace of AI development, particularly in the case of large language models (LLMs). Recent international reports on AI safety emphasize that emerging AI-related risks increasingly blur the boundary between security understood as protection against malicious threats and safety, which concerns ensuring that AI systems behave in a reliable, predictable, and intended manner[2, 6, 7].

In response to these challenges, regulatory and policy initiatives have gained significant momentum. A prominent example is the European Union Artificial Intelligence Act (AIA), which represents the first comprehensive legal framework worldwide aimed at regulating AI systems. The AIA classifies AI applications into four risk categories prohibited, high, limited, and minimal and associates each category with specific compliance and mitigation requirements. However, translating these regulatory principles into effective risk management practices poses considerable challenges, particularly for complex and adaptive AI systems such as LLMs[8-10].

One key difficulty lies in the fact that AI systems may exhibit emergent or unexpected capabilities beyond their original design intent, a phenomenon often referred to as function creep. While traditional risk analysis methodologies do consider emergent system behaviors, the scale, speed, and opacity of modern AI models introduce an additional layer of complexity that existing frameworks struggle to address adequately. Complementary efforts, such as the AI Risk Management Framework (AIRMF) proposed by the U.S. National Institute of Standards and Technology (NIST), further underline the growing international consensus on the need for structured approaches to AI risk governance[11-13].

Despite this increasing attention, many current approaches to AI risk management remain largely qualitative and frequently rely on risk matrices. Such tools are well known to suffer from conceptual and practical limitations, especially when applied to high-dimensional, adversarial, and rapidly evolving environments. These shortcomings are particularly problematic in the context of cybersecurity, where systems incorporating AI components face novel threats, including adversarial machine learning attacks and AI-enabled targeted intrusions[14].

Against this background, this paper proposes a strengthened framework for cybersecurity risk analysis tailored to systems that incorporate AI components, with particular emphasis on high-risk AI systems as defined by the EU AI Act, including those serving as product safety components. Building upon an existing adversarial risk analysis framework for cybersecurity, the proposed approach explicitly accounts for new AI-related impacts, AI-based assets, intelligent security and recovery controls, and emerging forms of targeted attacks. These elements are integrated into a unified analytical framework, which is subsequently illustrated and conceptually validated through a case study involving automated driving systems.

To clearly position the contribution of this work with respect to existing adversarial risk analysis frameworks, the proposed methodology combines inherited concepts, substantial methodological extensions, and several novel AI-specific components. The framework retains the fundamental adversarial risk analysis principles, including probabilistic attacker modeling, Monte Carlo-based risk estimation, attack propagation analysis, and cybersecurity portfolio optimization. These foundations are extended by explicitly representing AI-based assets, AI-driven security and recovery controls, AI-specific impact categories, and AI-enabled targeted attacks. Building upon these extensions, the principal novelty of this work lies in integrating all AI-specific elements into a unified quantitative cybersecurity risk analysis framework designed for high-risk AI systems, thereby providing a comprehensive methodology that supports both strategic cybersecurity investment decisions and compliance with emerging AI regulatory requirements. Section 2 explains these extensions in detail, while Section 3 presents the complete integrated framework.

2 Novel AI-related risk analysis issues and solutions

This section describes and resolves the issues suggested. When elsewhere available, we mainly summarize the literature with relevant pointers, particularly in Sections 2.3 and 2.4.

2.1 AI-related impacts

Traditional cybersecurity risk analysis has largely focused on technical impacts associated with confidentiality, integrity, and availability (CIA). While these dimensions remain fundamental, the increasing deployment of AI systems has brought additional categories of impacts to the forefront. In particular, concerns related to safety, bias, privacy, fairness, and accountability have emerged as critical issues, reflecting the broader societal and ethical implications of AI-enabled technologies. These concerns are explicitly recognized in several international guidelines and assessment tools, most notably in the Assessment List for Trustworthy Artificial Intelligence (ALTAI)[15-17].

As a result, existing cybersecurity objective catalogues require adaptation and, in some cases, extension to adequately capture these novel AI-related impacts. To address this need, we build upon the broader cybersecurity perspective introduced, which moves beyond purely technical considerations and incorporates business, organizational, and societal dimensions. In this approach, cybersecurity objectives are structured hierarchically in the form of a tree, encompassing impacts that can be directly quantified in monetary terms such as operational costs or equipment damage as well as impacts that are more difficult to monetize, including reputational harm, violations of fundamental rights, or loss of human life.

Our analysis is further informed by the AI Risk Management Framework (AIRMF), elements of which are also reflected in the EU AI Act, the OECD AI Recommendation, the ALTAI list, and reports such as those produced by the Center for Research in Foundation Models. Although these initiatives differ in terminology and structure, they converge on the concept of trustworthy AI. In this context, AI systems are expected to be valid, reliable, safe, secure, resilient, accountable, transparent, explainable, interpretable, privacy-enhancing, and fair, with each of these high-level characteristics comprising multiple subdimensions. For instance, validity entails requirements related to accuracy, robustness, and reliability[18-20].

Beyond defining desirable system properties, these frameworks also identify a range of potential adverse impacts associated with AI-related risks. Such impacts include, among others, harm to personal rights, discrimination, erosion of privacy, and threats to physical safety. The EU AI Act provides a more concrete articulation of these concerns by explicitly identifying prohibited practices and high-risk AI applications. An illustrative example of a forbidden functionality is the use of subliminal techniques that operate beyond an individual's consciousness with the intent of materially distorting behavior[11, 15, 21].

By mapping the trustworthiness characteristics and impact categories identified in these regulatory and governance frameworks onto an extended cybersecurity objective structure, it becomes possible to integrate AI-specific impacts into a unified risk analysis process. This integration enables a more comprehensive assessment of how AI systems may affect not only technical security outcomes but also organizational objectives, legal compliance, and societal values. Consequently, AI-related impacts can be systematically incorporated into preference models and risk evaluation procedures, providing a stronger foundation for informed decision-making in the cybersecurity risk management of systems with AI components[22-24].

2.2 AI-based assets

An increasing number of contemporary systems incorporate AI-based components that constitute valuable assets and, at the same time, introduce new vulnerabilities. Automated driving systems (ADSs) provide a paradigmatic example: perception and decision-making modules, typically based on machine learning techniques such as convolutional neural networks, are core elements of system functionality. These AI-based blocks are known to be susceptible to specialized attacks aimed at manipulating inputs or internal representations so as to induce erroneous outputs, potentially leading to severe safety and security consequences[25].

To properly capture these vulnerabilities within a cybersecurity risk analysis, organizations are modeled with a finer structural granularity than in traditional approaches. Rather than representing a system as a single monolithic entity, the architecture is decomposed into interconnected blocks (or components), which may correspond to hardware, software, or hybrid hardware software elements, some of

which are explicitly AI-based. These blocks are organized into levels according to criteria such as accessibility, security exposure, and criticality[26, 27].

Formally, consider a system structured into $k$levels. Blocks at level 1 are the most exposed and least critical, while blocks at higher levels are progressively more critical and typically less accessible from the outside. Attacks are assumed to enter the system through level-1 blocks and may then propagate to blocks at higher levels depending on the system architecture and the effectiveness of defensive controls. Let $P = \{p_i, p_{ij}, \dots, p_{12\cdots i_k}\}$ denote the set of entrance probabilities (EPs), where each term represents the probability that an external attack initially targets a specific combination of level-1 blocks. These probabilities satisfy the normalization condition $\sum_i p_i + \sum_{i<j} p_{ij} + \cdots + p_{12\cdots i_k} = 1$, ensuring that exactly one entry scenario occurs.

Once an attack has entered the system, its ability to compromise a given block depends on the effectiveness of the implemented defenses. This is modeled through probabilities of non-protection (PNPs). For external attacks against block B_i, let $Q_0 = [q_i]$ denote the probability that block $B_i$is *not* protected if attacked from outside. Consequently, $1 - q_i$represents the probability that the block successfully resists the attack.

Similarly, for attack propagation between blocks at different levels, let $Q_s = [q_{ij}]$ represent the probability that block $B_i$at level $s$is not protected against an attack transferred from block $B_j$at level $s$or $s-1$. The full set of non-protection probabilities is denoted by $Q = \{Q_0, Q_1, \dots, Q_k\}$.

These parameters explicitly account for the presence of AI-based assets and defenses. In particular, when a block corresponds to an AI component such as a perception model, content filter, or anomaly detector the associated PNP values depend on the robustness of the underlying learning algorithm and the defenses deployed against adversarial attacks. For AI-based safety or security components, empirical information from adversarial machine learning, such as security evaluation curves, can be used to estimate the probability of successful protection under attacks of varying type and intensity.

This block-based, probabilistic representation generalizes classical attack-graph models by allowing cycles, supporting heterogeneous components, and reflecting realistic cybersecurity architectures more closely. Importantly, it provides a structured way to integrate AI-based assets into the broader risk analysis framework, enabling the explicit modeling of how attacks may exploit AI vulnerabilities and propagate through complex systems.

2.3. AI-based security and recovery controls

Modern cybersecurity architectures increasingly rely on AI-based mechanisms not only as operational components but also as security and recovery controls. Examples include malware detectors, intrusion detection systems, anomaly detection modules, automated patching mechanisms, and AI-driven backup or recovery solutions. While these controls enhance protection capabilities, they also introduce new attack surfaces, as adversaries may specifically target or manipulate the AI components themselves through adversarial machine learning techniques[28, 29].

In particular, AI-based security controls can be deliberately deceived by carefully crafted inputs, leading to false negatives or incorrect system responses. Likewise, AI-driven recovery mechanisms may be compromised if the upstream detection components are successfully attacked. Consequently, the effectiveness of security and recovery controls must be explicitly modeled within the risk analysis framework, especially when such controls incorporate AI[30].

To this end, each defense mechanism is characterized through its impact on the probability of non-protection (PNP) of system blocks. Let $q \in [0,1]$denote the probability that a given block is *not protected* against a specific attack type under a particular defense formation. Accordingly, the probability of successful protection is given by $1 - q$.

When AI-based controls are deployed, the value of $q$depends on several factors, including the type of attack, its intensity, the learning model employed, and the robustness techniques implemented (e.g., adversarial training or ensemble defenses). These dependencies can be captured using security evaluation curves, which originate from the adversarial machine learning literature.

Formally, let αdenote the attack intensity and $d$the deployed defense. The probability of successful protection can be expressed as $1 - q = f_d(\alpha)$, where $f_d(\cdot)$is a defense-specific function mapping attack intensity to protection success probability. Consequently, the corresponding probability of non-protection is $q = 1 - f_d(\alpha)$.

These probabilities directly parameterize the PNP vectors introduced earlier. For external attacks against block B_i, the PNP is denoted by q_i, while for attack transfers from block $B_j$to block B_i, the PNP is denoted by q_{ij}. Both values depend on whether the relevant security or recovery controls AI-based or otherwise are implemented.

Let the vector $Q = Q_0, Q_1, \dots, Q_k$ collect all PNP parameters across levels, where $Q_0 = [q_i]$corresponds to external attacks and $Q_s = [q_{ij}]$corresponds to inter-block attack propagation at level s. Different security portfolios yield different realizations of Q, thereby altering the system's overall risk profile.

Importantly, AI-based controls may serve dual roles. On the one hand, they act as protective mechanisms intended to reduce attack success probabilities; on the other hand, they may themselves be targeted by adversaries. This duality requires modeling assumptions that explicitly recognize the possibility of control failure induced by adversarial attacks, rather than treating defenses as static or perfectly reliable.

The parameters $q$can be estimated using a combination of empirical data, expert judgment, and experimental evaluations. In the case of AI-based defenses, adversarial robustness benchmarks and security curves provide a principled way to quantify defensive performance under different threat scenarios. These estimates can be updated as new evidence becomes available, allowing the framework to support adaptive and forward-looking risk assessments.

By explicitly incorporating AI-based security and recovery controls into the probabilistic structure of the model, the framework enables a more realistic representation of modern cybersecurity systems. This approach captures both the benefits and limitations of intelligent defenses and provides a coherent foundation for evaluating alternative protection portfolios in environments where AI plays a central role.

2.3 AI-Based Targeted Attacks

Similar to other areas of cybersecurity, attacks against AI-enabled systems may be untargeted (opportunistic) or targeted (strategic). In the AI domain, targeted attacks are particularly relevant because adversaries can themselves exploit AI technologies. Examples include the use of adversarial perturbed inputs to deceive machine learning models, or the injection of malicious data that exploits *function creep* in deployed AI systems.

The increasing availability of AI-powered Crime-as-a-Service platforms has significantly lowered the barrier for launching sophisticated attacks. From a modeling standpoint, it is not essential whether these attacks are manually crafted or automated using AI; what matters is whether attackers possess sufficient technical capabilities or resources to deploy advanced AI-based attack methods[31].

As discussed in Section 2.2, two key elements must be modeled for each attack type:

1. The entry probabilities (EPs) into different system blocks.
2. The stochastic process governing the arrival of attacks.

When attacks are targeted, estimating entry probabilities becomes more complex, because attackers behave strategically. To address this issue, the framework adopts an Adversarial Risk Analysis (ARA) approach, which explicitly models attacker decision-making under uncertainty. The ARA process consists of four main steps:

1. **Defender problem formulation**: The defender selects a cybersecurity portfolio $c$to minimize the expected impact of attacks. Uncertainty includes whether the system will be targeted, which attack type will be used, and which entry point will be chosen, given the deployed defenses.
2. **Attacker problem formulation**: The defender models the attacker's beliefs and preferences using probability distributions, reflecting incomplete information about attacker objectives, costs, and success probabilities.
3. **Simulation of attacker decisions**: By sampling from the attacker's uncertain utilities and probabilities, the defender simulates the attacker's optimal choices. Repeating this process yields estimates of attack probabilities and preferred entry points.
4. **Defensive optimization**: These estimated attack probabilities are incorporated back into the defender's optimization problem to select the most effective cybersecurity portfolio.

Attacker Action Space

Assume the attacker can choose from the following set of actions $\mathcal{A}$:

1. **Targeting the defender's system** $a_1^{(j,k)}$ where:
   - $j \in \{1, \dots, J\}$denotes the attack type,
   - $k \in \{1, \dots, K\}$denotes the entry point (or combination of entry points).
2. **Targeting alternative systems** $a_i^j, i \in \{2, \dots, n\}$

representing attacks of type $j$against other potential targets.

The attacker is assumed to select one attack action at a time.

Attacker Expected Utility

Let $Y \in \{0,1\}$denote the outcome of an attack, where:

- $Y = 1$: attack succeeds,
- $Y = 0$: attack fails.

Given portfolio c, the attacker's expected utility for attacking the defender's system via entry $k$with attack type $j$is:

$$\psi_A\left(a_1^{(j,k)}, c\right) = p_A(Y = 1 \mid a_1^{(j,k)}, c)\, u_A(Y = 1, a_1^{(j,k)}, c) + p_A(Y = 0 \mid a_1^{(j,k)}, c)\, u_A(Y = 0, a_1^{(j,k)}, c)$$

Because the defender does not know the attacker's exact utilities and beliefs, these quantities are modeled as random variables:

$$\tilde{\psi}_A\left(a_1^{(j,k)}, c\right) = P_A(Y = 1 \mid a_1^{(j,k)}, c)\, U_A(Y = 1, a_1^{(j,k)}, c) + P_A(Y = 0 \mid a_1^{(j,k)}, c)\, U_A(Y = 0, a_1^{(j,k)}, c)$$

Similarly, for attacking another system i:

$$\tilde{\psi}_A(a_i^j) = P_A(Y = 1 \mid a_i^j)\, U_A(Y = 1, a_i^j) + P_A(Y = 0 \mid a_i^j)\, U_A(Y = 0, a_i^j)$$

Changes in the defenses of competing systems are not modeled due to limited information.

Attacker Optimal Decision

Given portfolio c, the attacker selects the action that maximizes random expected utility:

$$\zeta(c) = \arg\max_{x \in \mathcal{A}} \tilde{\psi}_A(x)$$

This optimization is approximated via Monte Carlo simulation. At iteration v:

$$\zeta_v(c) = \arg\max_{x \in \mathcal{A}} \tilde{\psi}_A^{(v)}(x)$$

Using Vsimulations, the probability that the attacker targets the defender's system with attack type $j$is estimated as:

$$\tau_j^1(c) = \mathbb{P}_D(A = a_{j,\cdot}^1 \mid c) = \frac{|\{\zeta_v(c) = a_{j,\cdot}^1\}|}{V}$$

Similarly, probabilities of attacking other systems are computed.

Entry Point Probabilities

Conditional on attack type j, the probabilities of targeting each entry point are estimated as: $\gamma_j^1(c) = (\gamma_{j,1}^1, \dots, \gamma_{j,K}^1)$ where:

$$\gamma_j^1(c) = \left(|\{\zeta_v = a_{j,1}^1\}| + 1,\ \dots,\ |\{\zeta_v = a_{j,K}^1\}| + 1\right)$$

The Laplace smoothing term $+1$prevents zero probabilities. These vectors parameterize a Dirichlet distribution, from which entry probabilities (EPs) are sampled for use in attack simulations.

Algorithms 1–5 jointly describe the computational workflow of the proposed framework, covering attack propagation, impact assessment, cyberattack simulation, expected utility estimation, and cybersecurity portfolio optimization.

Algorithm 1. Attack Propagation Simulation in an AI-Enabled System

Input:

- $G = (V, E)$: Hierarchical system graph
- EP: Attack entry probability vector
- PNP: Protection failure probabilities
- T_{max}: Maximum number of attack transitions

Output:

- C: Set of compromised blocks

1. Initialize $C = \emptyset$.
2. Sample the attack entry block according to EP.
3. Add the selected block to C.
4. Set Frontier = entry block.
5. Set $t = 0$.
6. While (Frontier is not empty) and ($t < T_{max}$) do

Set New Frontier = $\emptyset$.

For each block $b_i$in Frontier:

- For each neighboring block b_j:
  - If $b_j \notin C$:
    - Generate a random number $u \sim U(0,1)$.
    - If $u \leq PNP(b_i, b_j)$:

- Mark b_j as compromised.
- Add b_j to C.
- Add b_j to NewFrontier.
- End If
- End If
- End For

End For
Set Frontier = New Frontier.
Set $t = t + 1$.
7. End While
8. Return C.

Key Takeaway

This ARA-based modeling approach allows the defender to:

- Anticipate whether attackers will target the system at all,
- Predict which attack types are most likely,
- Estimate which system entry points are most vulnerable,
- And incorporate all of this strategically into cybersecurity portfolio optimization.

2.4 Summary of Framework Contributions and Novel Components

The proposed framework builds upon the adversarial risk analysis (ARA) methodology for cybersecurity while introducing several extensions specifically designed for AI-enabled systems. To clearly distinguish between inherited concepts, modified components, and original contributions, the framework can be summarized as follows.

Inherited components. The proposed methodology retains the fundamental principles of adversarial risk analysis, including probabilistic modeling of attacker behavior, Monte Carlo-based risk estimation, attack propagation analysis, and cybersecurity portfolio optimization. The general concepts of attack entry probabilities, protection failure probabilities, and utility-based decision analysis are also inherited from the original ARA framework.

Modified components. Several elements of the original framework have been substantially extended to address the characteristics of AI-enabled systems. First, the system architecture is refined to explicitly represent AI-based components as individual blocks within the hierarchical system model. Second, traditional cybersecurity impacts are expanded to include AI-related concerns such as safety, fairness, accountability, privacy, and regulatory compliance. Third, security and recovery controls are generalized to incorporate AI-driven defensive mechanisms whose effectiveness depends on adversarial robustness. Finally, attacker behavior is extended to account for AI-enabled targeted attacks and adversarial machine learning techniques.

Novel components. The principal contributions of this work are the introduction of an integrated quantitative framework that explicitly incorporates AI-specific cybersecurity characteristics into adversarial risk analysis. These contributions include: (i) explicit modeling of AI-related impact categories within the cybersecurity objective hierarchy; (ii) representation of AI-based assets and learning-enabled components within the attack propagation model; (iii) probabilistic integration of AI-based security and recovery controls using adversarial robustness information; (iv) incorporation of AI-enabled targeted attacks into attacker decision modeling; and (v) a unified simulation framework that combines these elements into a quantitative risk assessment and cybersecurity investment methodology applicable to high-risk AI systems.

This separation between inherited, extended, and novel elements clarifies the methodological contribution of the proposed framework and highlights how it advances existing adversarial risk analysis methods to address the cybersecurity challenges posed by modern AI-enabled systems.

3 The Framework

This section integrates the solutions introduced earlier into a comprehensive cybersecurity risk analysis framework tailored for systems that include AI components. The framework is organized as a five-stage pipeline, covering attack propagation, impact simulation, attack generation, risk assessment, and risk management.

3.1 Simulation of Attack Propagation

Based on the block-level architecture defined in Section 2.2, an algorithm is introduced to simulate how an attack propagates through a system's components.

Each system block B_i is associated with a binary indicator:

$$I_i = \begin{cases} 1 & \text{if block } B_i \text{ is successfully compromised} \\ 0 & \text{otherwise} \end{cases}$$

The simulation accounts for:

- Initial attack entry points,
- Protection failure probabilities (PNPs),
- Attack transfers between blocks,
- Possible cycles in the system graph.

Let N' denote the maximum number of attack transitions allowed before detection or termination.

3.2 Impact Simulation and Aggregation

Once the compromised blocks are identified, the next step is to simulate the impacts caused by the attack.

Two types of impacts are considered:

- Local impacts: depend on individual blocks (e.g., component damage).
- Global impacts: affect the entire system (e.g., legal liability, reputation).

Let: L be the number of local impacts, R be the total number of impacts.

Local impacts are aggregated using block-specific aggregation functions $b_j(\cdot)$, while global impacts are directly sampled. All impacts are then combined using a global aggregation function $g(\cdot)$.

3.3 General Attack Simulation Scheme

This stage integrates attack arrival processes, attacker decision models, and impact simulation. Each attack type has its own stochastic arrival process Φ. Attacks may be targeted (modeled via ARA) or untargeted.

Algorithm 2. Impact Simulation and Loss Evaluation

Input:

- C: Set of compromised system blocks
- I_L: Local impact distributions
- I_G: Global impact distributions
- W: Impact weights
- N: Number of Monte Carlo simulations

Output:

- L: Estimated total loss

1. Initialize the total loss $L \leftarrow 0$.
2. Set the simulation counter $k \leftarrow 1$.
3. While $k \leq N$ **do**

Initialize the cumulative local impact $L_{local} \leftarrow 0$.

For each compromised block $b_i \in C$:

- Sample the local impact value from the corresponding probability distribution.
- Multiply the sampled impact by its associated weight.
- Update the cumulative local impact.

End For

Sample the global system impacts from their corresponding probability distributions.

Compute the aggregated system loss by combining local and global impacts.

Store the simulated loss value.

Set $k \leftarrow k + 1$.

4. End While
5. Compute the expected loss as the average of all simulated losses.
6. Return the estimated total loss L.

Algorithm 2 estimates the consequences of a successful cyberattack by combining local impacts associated with compromised components and global impacts affecting the overall AI-enabled system. The impact values are generated through Monte Carlo sampling from the corresponding probability distributions and aggregated to obtain the total system loss. The resulting loss estimates are subsequently used for adversarial risk analysis and cybersecurity portfolio optimization.

Algorithm 3. Cyberattack Simulation

Input:

- S: System model
- EP: Attack entry probability vector
- PNP: Protection failure probabilities
- A: Set of possible attack types
- N: Number of simulations runs

Output:

- $\mathcal{L}$: Simulated loss distribution

1. Initialize the loss set $\mathcal{L} \leftarrow \emptyset$.
2. Set the simulation counter $i \leftarrow 1$.
3. While $i \leq N$ **do**

Randomly select an attack type from A.

Generate an attack entry point according to the entry probabilities EP.

Execute Algorithm 1 to determine the compromised system components.

Execute Algorithm 2 to evaluate the corresponding system loss.

Store the obtained loss value in $\mathcal{L}$.

Set $i \leftarrow i + 1$.

4. End While
5. Compute the empirical loss distribution from $\mathcal{L}$.
6. Return $\mathcal{L}$.

Algorithm 3 performs the overall cyberattack simulation process. For each Monte Carlo replication, an attack type and an attack entry point are generated according to the corresponding probability models. The attack propagation is simulated using Algorithm 1, while the resulting consequences are evaluated through Algorithm 2. Repeating this process produces the empirical loss distribution required for the subsequent adversarial risk analysis.

3.4 Risk Assessment

Using the Monte Carlo loss samples, the system loss distribution is estimated. The distribution is modeled as a mixed distribution: $L \sim s \cdot \delta_0 + (1 - s) \cdot f(l)$ where:

- s is the probability of zero loss,
- δ_0 is a point mass at zero,
- $f(l)$ is a continuous density, typically approximated by a **Gamma mixture**.

From this distribution, standard risk measures are computed:

- Expected loss,
- Value at Risk (VaR),
- Conditional Value at Risk (CVaR),
- Loss exceedance curves.

If the assessed risk exceeds acceptable levels, the process proceeds to risk management.

3.5 Risk Management

Risk management focuses on selecting an optimal cybersecurity portfolio.

Let: $c = (c_1, \dots, c_m), c_i \in \{0,1,2\}$ where:

- 0: not implemented,
- 1: candidate control,
- 2: already implemented.

The optimization problem is: $\max_c \ \mathbb{E}[u(c)]$ subject to:

$$\begin{cases} c_i = 2 & i \le r \\ c_i = 1 & r < i \le s \\ h(c) \le 0 \end{cases}$$

Utility Function

A CARA utility function is adopted: $u(c) = 1 - \exp\left(\rho(l(c) + \text{cost}(c))\right)$ where:

- ρ is the risk aversion coefficient,
- $l(c)$ is the stochastic loss,
- $\text{cost}(c)$ includes implementation and maintenance costs.

Loss Approximation Model

Losses are approximated as: $L(c) \sim s(c)\, \delta_0 + (1 - s(c))\, \Gamma(a, t(c))$ with: $s(c) = 1 - s_0 \exp\left(-\sum_{i=1}^{m} \alpha_i \ \text{sign}(c_i)\right),\ t(c) = t_0 - \sum_{i=1}^{m} \beta_i \ \text{sign}(c_i)$

Monte Carlo simulation is used to estimate expected utility:

$$\hat{u}(c) = \frac{1}{M} \sum_{i=1}^{M} (1 - \exp(\rho \cdot \text{cost}_i))$$

where each cost_i is generated from the mixture model above.

Algorithm 4. Expected Utility Estimation

Input:

- $\mathcal{P}$: Candidate cybersecurity portfolio
- $\mathcal{L}$: Loss distribution obtained from Algorithm 3
- $U(\cdot)$: Utility function
- N: Number of Monte Carlo simulations

Output:

- $EU(\mathcal{P})$: Expected utility of the candidate portfolio

1. Initialize the cumulative utility $TU \leftarrow 0$.
2. Set the simulation counter $i \leftarrow 1$.
3. While $i \le N$ **do**

Apply the candidate cybersecurity portfolio $\mathcal{P}$.

Execute Algorithm 3 to simulate one cyberattack scenario.

Obtain the simulated system loss L_i.

Compute the utility value $U(L_i)$.

Update the cumulative utility:

$TU \leftarrow TU + U(L_i)$.

Set $i \leftarrow i + 1$.

4. End While
5. Compute the expected utility:

$EU(\mathcal{P}) = \frac{TU}{N}$.

6. Return $EU(\mathcal{P})$.

Algorithm 4 estimates the expected utility associated with a candidate cybersecurity portfolio. For each Monte Carlo replication, the portfolio is applied to the system, a cyberattack scenario is simulated using Algorithm 3, and the corresponding utility is evaluated. The average utility over all simulations provides the expected utility, which serves as the performance measure for comparing alternative cybersecurity investment strategies.

Algorithm 5. Cybersecurity Portfolio Optimization

Input:

- $\mathcal{P}$: Set of feasible cybersecurity portfolios
- B: Available cybersecurity budget
- $C(\mathcal{P})$: Portfolio implementation cost
- $U(\cdot)$: Utility function
- N: Number of Monte Carlo simulations

Output:

- $\mathcal{P}^*$: Optimal cybersecurity portfolio
- EU^*: Maximum expected utility

1. Initialize the best portfolio $\mathcal{P}^* \leftarrow \emptyset$.
2. Initialize the maximum expected utility $EU^* \leftarrow -\infty$.
3. For each candidate portfolio $\mathcal{P}_i$ **do**

Compute the implementation cost $C(\mathcal{P}_i)$.

If $C(\mathcal{P}_i) \leq B$ then

- Execute Algorithm 4 to estimate the expected utility $EU(\mathcal{P}_i)$.
- **If** $EU(\mathcal{P}_i) > EU^*$ **then**
  - Update $EU^* \leftarrow EU(\mathcal{P}_i)$.
  - Update $\mathcal{P}^* \leftarrow \mathcal{P}_i$.
- End If

End If

4. End For
5. Return the optimal portfolio $\mathcal{P}^*$ and the corresponding expected utility EU^*.

Algorithm 5 identifies the optimal cybersecurity investment portfolio under a predefined budget constraint. Each feasible portfolio is evaluated by estimating its expected utility through Algorithm 4, which internally incorporates cyberattack simulation, attack propagation, and impact assessment. The portfolio achieving the highest expected utility while satisfying the budget limitation is selected as the optimal cybersecurity strategy.

Summary of Section 3

This framework:

- Explicitly models AI-specific attack dynamics,
- Captures attacker strategic behavior,
- Simulates attack propagation at component level,
- Produces realistic loss distributions,
- Supports rational, risk-aware cybersecurity investment decisions.

4 Case Study

This section illustrates the practical application of the proposed cybersecurity risk analysis framework through a simplified but realistic case study. The example is designed to be sufficiently rich to demonstrate all modeling stages, while remaining tractable for interpretation. The case also highlights both the benefits and limitations of the framework, particularly in terms of modeling effort and computational cost.

The study focuses on an Automated Driving System (ADS) fleet operator that provides vehicle rental services. Two competing companies operate in the same market. Given the heavy reliance of ADSs on AI components, improving cybersecurity is a strategic priority.

4.1 Model Setup (System Architecture)

4.1.1 Parameter Sources and Probability Elicitation

The numerical parameters adopted in the case study were selected to provide a representative evaluation of the proposed framework rather than to reproduce the cybersecurity profile of a particular industrial deployment. Since publicly available datasets reporting detailed attack probabilities, attack propagation parameters, and AI-specific defense effectiveness for automated driving systems remain scarce, the model parameters were obtained through a structured parameter elicitation process.

Specifically, attack entry probabilities, protection failure probabilities (PNPs), and impact distributions were determined by combining information available in the cybersecurity and adversarial machine learning literature with engineering judgment regarding the architecture of AI-enabled automated driving systems. The selected values were verified to satisfy probability consistency constraints and to remain within ranges commonly reported for comparable cybersecurity studies.

The parameter estimation process consisted of four consecutive steps. First, the system architecture and attack scenarios were defined based on the ADS configuration. Second, attack probabilities and protection parameters were

assigned using published evidence and expert assessment. Third, probability values were adjusted to ensure internal consistency across attack propagation paths. Finally, Monte Carlo simulations were performed to verify that the resulting loss distributions remained realistic and numerically stable before conducting the risk analysis.

The ADS architecture follows the block structure introduced earlier and consists of:

- Perception (P) level 1,
- Localization (L) level 1,
- Decision-Making (DM) level 2.

Only level-1 blocks are directly accessible from external attacks, while higher-level blocks can be compromised through propagation. Three impact dimensions are considered over a one-year planning horizon:

1. Financial losses: Including legal liability, data breaches, and regulatory penalties: (measured in thousands of euros).
2. Equipment damage: Cost of repairing or replacing hardware/software components: (thousands of euros).
3. Downtime: ADS unavailability duration: (measured in hours).

Untargeted Threats

To simplify analysis, two untargeted threats are considered:

1. Denial of Service (DoS): Attacks on traffic infrastructure or vehicle interfaces to block access or extort users.
2. Supply Chain Threats (SCTs): Malicious software or firmware updates that degrade or disable ADS functionality.

Targeted Threats and Attackers

Two attacker profiles are modeled:

- Cyberterrorist group (Cy) Motivated by political impact and notoriety.
- Cybercriminal group (Cr) Motivated by economic gain.

Table 1 provides a summary of attack categories and the threat actors associated with each. For modeling purposes, attackers are assumed to launch only a single type of attack at any given time. This restriction is applicable only to the cyberterrorist group, which is capable of executing both attack types. The impacts resulting from the different threats are presented in Table 2.

Table 1. Attacks available to malicious actors

Attacks/attacker	C_y	C_r
AML_at	X	X
wir_jam	X	

Table 2. Impacts deemed relevant for targeted and untargeted threats

Attacks /impacts	**Financial**	**Equipment damage**	**Downtime**
$AML\_at\ (Cy)$	X	X	X
$AML\_at\ (Cr)$	X	X	
$Wireless\ jam.$	X	X	
DoS	X		
$SCTs$	X	X	

Table 3. Defense effectiveness

Attacks/defenses	**FwGw**	**AML**	**PmVs'**
$Adversarial\ attack\ X$		X	
$Wireless\ jamming$	X		X
DoS	X		X
$SCTs$	X	X	X

Table 4. Impacts mitigated by each cyber insurance product

Financial	***Equipment damage***	**Downtime**
A	X	
B	X	X

Potential countermeasures under consideration, but not yet deployed within the ADS, include a firewall and internet gateway (FwGw), an enhanced anti-money laundering module (AML), and a combination of patch management, intrusion detection, and vulnerability scanning mechanisms (PmVs). The effectiveness of these defenses in mitigating attacks either by reducing their likelihood of occurrence or minimizing their impact when successful is summarized in Table 3. Additionally, two cyber insurance plans are evaluated: a basic policy (A), which covers physical damage to vehicle equipment, and an advanced policy (B), which extends coverage to include losses resulting from operational downtime, as detailed in Table 4.

The following targeted attacks are considered:

1. Adversarial ML attacks (AML_at) Manipulation of data inputs to deceive ADS machine learning models.

2. Wireless jamming (wir_jam) Disruption of GPS or communication signals affecting both control and ML subsystems.

Attack availability is summarized conceptually as: Cy → {AML_at,wir_jam},Cr → {AML_at}
Only one attack can be executed at any given time by an attacker.

Defensive Controls, Three defensive measures are considered:

1. Firewall and Gate way ($FwGw$)
2. Robust AML defense module (AML)
3. Patch management, IDS, and vulnerability scanner (PmVs)

Their effectiveness against attacks is modeled by modifying attack probabilities and/or impact severity.

Cyber Insurance, two insurance products are available:

- Insurance A: covers equipment damage.
- Insurance B: covers equipment damage and downtime.

At least Insurance A is mandatory due to regulatory requirements. The limitations include;

- Maximum cybersecurity budget: $C_{\max} = 3400$ euros
- Insurance A must be included in all feasible portfolios.

4.1.2 Model Calibration

The numerical parameters were calibrated iteratively to ensure that the simulated behavior of the system remained consistent with the expected cybersecurity characteristics of AI-enabled automated driving systems reported in the literature. Calibration focused on maintaining realistic relationships between attack frequencies, attack propagation probabilities, defense effectiveness, and resulting loss distributions. Rather than reproducing a specific industrial deployment, the calibration aimed to generate a representative benchmark scenario suitable for evaluating the proposed quantitative framework.

4.1.3 Computational Implementation

All numerical experiments were performed using Monte Carlo simulation. Unless otherwise stated, each experiment consisted of 10,000 independent simulation replications, providing stable estimates of the expected loss and expected utility. Preliminary experiments indicated that increasing the number of simulations beyond this value produced only negligible changes in the estimated results.

Table 5 summarizes the principal computational settings adopted throughout the numerical experiments, allowing independent reproduction of the proposed framework.

Table 5. Computational Settings Used in the Case Study

Setting	Value
Case study	AI-enabled Automated Driving System (ADS)
Simulation approach	Monte Carlo simulation
Number of Monte Carlo replications	10,000
Attack propagation model	Algorithm 1
Impact evaluation	Algorithm 2
Cyberattack simulation	Algorithm 3
Portfolio evaluation	Algorithm 4
Portfolio optimization	Algorithm 5
Optimization method	Exhaustive evaluation of feasible portfolios
Programming environment	MATLAB R2024a
Random number generator	MATLAB default (Mersenne Twister)
Output measures	Expected Loss, Expected Utility, Optimal Portfolio

All simulations were implemented in MATLAB using the proposed computational framework. Monte Carlo simulation was employed to evaluate the probabilistic behavior of cyberattacks and to estimate the expected loss and expected utility associated with alternative cybersecurity portfolios.

4.2 Risk Analysis Results

Initial Risk Assessment

The baseline formation includes only mandatory insurance A and no additional defenses.

Using Monte Carlo simulation with: $M = 10{,}000$ loss samples are generated, and The loss distribution exhibits:

- Zero-loss probability: $s = 0$
- Positive losses approximated by a Gamma distribution:
  $L \sim \Gamma(a, p), a = 5871.48, p = 0.00235$

Risk measures:

- 95% Value at Risk (VaR): $\mathrm{VaR}_{0.95} = 1.32$ M€
- 95% Conditional Value at Risk (CVaR): $\mathrm{CVaR}_{0.95} = 1.498$ M€

This risk level is deemed unacceptable, triggering the risk management phase. Risk Management and Portfolio

Optimization Out of 16 possible portfolios, 12 satisfy budget and compliance constraints. Expected utility is estimated using a CARA utility function:

$$u(c) = 1 - \exp\left(\rho(l(c) + \text{cost}(c))\right)$$

Monte Carlo simulation identifies the optimal portfolio as: Insurance A + AML + PmVs

This portfolio is not the most expensive, yet provides the best trade-off between cost and risk reduction.

Risk Reduction Achieved, For the optimal portfolio:

- Probability of zero loss: $s = 0.174$
- Positive losses: $L \sim \Gamma(a, p), a = 192.76, p = 0.05189$

Risk measures:

- 95% VaR: 59,520 €
- 95% CVaR: 72,756 €

Compared with the baseline setup, the proposed approach achieves a significant reduction in tail risk. Table 6 provides an overview of the three top-performing portfolios as well as the initial solution, including their expected loss, deployment cost, and expected utility metrics.

Table 6. Three best portfolios and initial formation

Attacks/defenses	FwGw	AML	PmVs
Adversarial attack X		X	
Wireless jamming	X		X
DoS	X		X
SCTs	X	X	X

Table 7. Probability of perpetrating an attack on our company or attacking another one by the cyberterrorists

Attacks/defenses	FwGw	AML	PmVs
Adversarial attack X		X	
SCTs	X	X	X

Table 8. Γ vector of each targeted attack for initial and optimal portfolio configuration

AML_at (Cy)			
	P	L	P, L
Initial	0.33	0.34	0.33
Optimal	0.35	0.30	0.35
Wir jam (Cy)			
	P	L	P, L
Initial	0.32	0.34	0.34
Optimal	0.34	0.33	0.33

The proposed framework also delivers additional security insights within the given context. Based on the methodology described in Section 2.4, Table 7 reports the probabilities associated with different cyberterrorist attacks under both the initial and optimized portfolios. The results highlight the deterrence capability of the optimized portfolio, as a substantial portion of the attack likelihood is redirected toward competing targets, while protection against AML-related attacks becomes nearly complete. Furthermore, Table 8 presents a more detailed analysis of attack entry points for the cybercriminal threat group. In contrast to the cyberterrorist scenario, only negligible differences are observed between the initial and optimized portfolios, suggesting that all entry points exhibit a similar level of vulnerability to both attack types.

For the cyberterrorist group:

$$\mathbb{P}(\text{AML\_at} \mid \text{initial}) = 0.63$$
$$\mathbb{P}(\text{AML\_at} \mid \text{optimal}) = 0.01$$

Most attack probability mass shifts toward other competing companies, demonstrating a deterrence effect. The estimated entry point probability vectors for targeted attacks are approximately uniform:

$$\gamma = (P, L, P + L) \approx (0.33, 0.33, 0.34)$$

These probabilities exhibit minimal variation across different portfolio configurations, suggesting that the implemented defenses mainly influence the practicality and success likelihood of attacks rather than the attackers' choice of entry point. To further evaluate the robustness of the proposed approach, a comprehensive sensitivity analysis was conducted on several key parameters. The first analysis examined the effect of budget constraints to support budget allocation and negotiation decisions. As shown in Table 9, optimal portfolio compositions were determined for budget levels ranging from 600 to 4500. The results indicate a substantial decrease in expected loss with each incremental budget increase, except for the final budget level, where the portfolio transitions to insurance option B. Since insurance B is selected only under the highest budget scenarios, an additional analysis was performed to assess the effect of reducing its premium by up to 20% while maintaining the same coverage benefits. The corresponding optimal portfolios for different discount levels are presented in Table 10.

Table 9. Optimal portfolio for different values of maximum budget

Portfolio	Expected loss	Expected utility	Budget
A	1,016,077	−0.107	[600,700]
A, AML	814,828	−0.085	[800,1700]
$A, FwGw, AML$	48,185	−0.005	[1800,2200]
$A, AML, PmVs$	22,834	−0.002	[2300,4200]
$A, FwGw, AML, PmVs$	13,337	−0.0017	[3500,4200]
$B, FwGw, AML, PmVs$	11,743	−0.0016	[4300,4500]

Table 10. Optimal portfolio for several price reductions in insurance B

Price reduction	*Portfolio*	*Expected loss*	*Cost*	*Expected utility*
20%	$B, AML, PmVs$	20,338.85	3171	-0.0024
10%	$B, AML, PmVs$	20,441.08	3311	-0.0024
4%	$B, AML, PmVs$	20,508.96	3395	-0.0024
3%	$A, AML, PmVs$	22,491.14	2301	-0.0025

Sensitivity Analysis, Varying the maximum budget produces the following optimal strategies:

- Low budget: insurance only
- Medium budget: insurance + AML
- Higher budget: insurance + AML + PmVs
- Highest budget: full defense stack

This confirms robustness of the optimal portfolio under reasonable budget variations.

To evaluate the robustness of the proposed methodology with respect to parameter estimation, additional sensitivity analyses were performed by varying the principal probabilistic parameters within plausible ranges. In particular, attack frequencies and protection failure probabilities (PNPs) were independently varied by ±20% while keeping all other parameters unchanged. The resulting analyses showed only moderate variations in the estimated risk measures, whereas the ranking of the optimal cybersecurity portfolios remained unchanged. These results indicate that the proposed framework is reasonably robust against moderate uncertainty in parameter estimation and does not rely on a narrow set of numerical assumptions.

4.3 Uncertainty Analysis

Uncertainty is an inherent characteristic of cybersecurity risk analysis, particularly for AI-enabled systems where empirical evidence is still limited. In the proposed framework, uncertainty affects attack occurrence rates, entry probabilities, protection failure probabilities, and impact distributions. These uncertainties are explicitly propagated through Monte Carlo simulation, allowing the estimated loss distributions to reflect both stochastic attack behavior and uncertainty in the underlying model parameters. Consequently, the reported risk measures should be interpreted as probabilistic estimates rather than deterministic values.

5 Discussion and open issues

Motivated by the ‘EU AI Act, and related international policy developments, we suggested how to deal with novel risk analysis issues emerging in systems with AI components. Potential benefits were illustrated with an ADS case. Beyond as a risk analysis framework, its risk assessment stage could be used to determine where at the four-tier EU AI Act risk ladder a system is, to understand whether it is required to adopt and implement additional security controls. This flexibility contributes to define a robust framework capable of integrating advanced risk analysis and protection methods for AI applications but also to support certification programs as demanded in the EU Cybersecurity and Cyber-resilience Acts (Cihon, 2019)’ usable as well in a proactive manner to check the impact of protection measures through what-if analyses.

Several measures ‘emerged recently to treat the safety, ethical, bias, privacy, and fairness threats that are becoming of major concern in the AI domain (Ala-Pietilä et al., 2020), like mechanisms to enforce antidiscrimination policies. Although not exactly cybersecurity controls, they can be inte-grated into the framework, aggregating them as procedural, technical, or physical controls against unwarranted actions, to be used in a broader risk analysis process. Besides, the inclusion of security and other risk-related objectives (e.g., privacy or fairness) is essential when developing trustworthy AI systems. Our framework could be used to address the design of AI systems, thus embedding a security-by-design approach’ matching the AI and cybersecurity securitization and russification efforts.

As the case study showed, our approach is technical and demands intensive modeling work. However, the values at stake are so important that the extra effort should be worth it. To facilitate its implementation, a system could be developed, possibly adopting the ENISA terminologies, extended with the novel AI ingredients presented here. Importantly, many of these will be common across systems with similar distributions for blocks with the same structure and formation, enabling the creation of reusable templates for systems with such similarities. This would alleviate elicitation tasks and allow for the development of block templates, security curves, safety components, impacts, preference models, and distributions.

The framework provides a strategic approach to cybersecurity risk analysis for systems with AI components with advantages over standards using risk matrices. Recent interesting proposals [29] facilitate operational *sense-and-*

respond approaches to AI-based cybersecurity mostly based on partially observable Markov decision processes which complement our proposal: on the one hand, the corresponding systems could be included as safety components within the considered security portfolio; on the other, the ingredients presented, mainly explicit utility functions and ARA calculations to estimate attack and defense probabilities, could be used to improve their modeling.

Ethical Considerations

All procedures performed in this study were under the ethical standards.

Acknowledgments

Authors thank all who helped us through this study.

Conflict of Interest

The authors report no conflict of interest.

Funding

According to the authors, this article has no financial support.

References

[1] J. Perez-Cerrolaza *et al.*, "Artificial intelligence for safety-critical systems in industrial and transportation domains: A survey," *ACM Computing Surveys,* vol. 56, no. 7, pp. 1-40, 2024.
[2] J. M. Camacho, A. Couce-Vieira, D. Arroyo, and D. R. Insua, "A cybersecurity risk analysis framework for systems with artificial intelligence components," *International Transactions in Operational Research,* vol. 33, no. 2, pp. 798-825, 2026.
[3] B. W. Wirtz, J. C. Weyerer, and I. Kehl, "Governance of artificial intelligence: A risk and guideline-based integrative framework," *Government information quarterly,* vol. 39, no. 4, p. 101685, 2022.
[4] M. Doumpos, C. Zopounidis, D. Gounopoulos, E. Platanakis, and W. Zhang, "Operational research and artificial intelligence methods in banking," *European journal of operational research,* vol. 306, no. 1, pp. 1-16, 2023.
[5] K. Zhang *et al.*, "Artificial intelligence in drug development," *Nature medicine,* vol. 31, no. 1, pp. 45-59, 2025.
[6] P. Radanliev, D. De Roure, C. Maple, J. R. Nurse, R. Nicolescu, and U. Ani, "AI security and cyber risk in IoT systems," *Frontiers in big data,* vol. 7, p. 1402745, 2024.
[7] R. Dobbe, "AI Safety is Stuck in Technical Terms--A System Safety Response to the International AI Safety Report," *arXiv preprint arXiv:2503.04743,* 2025.
[8] P. Hacker, A. Engel, and M. Mauer, "Regulating ChatGPT and other large generative AI models," in *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, 2023, pp. 1112-1123.
[9] D. Rios Insua, A. Couce-Vieira, J. A. Rubio, W. Pieters, K. Labunets, and D. G. Rasines, "An adversarial risk analysis framework for cybersecurity," *Risk Analysis,* vol. 41, no. 1, pp. 16-36, 2021.
[10] N. Kaloudi and J. Li, "The ai-based cyber threat landscape: A survey," *ACM Computing Surveys (CSUR),* vol. 53, no. 1, pp. 1-34, 2020, doi: https://doi.org/10.1145/3372823.
[11] M. Yazdi, S. Adumene, D. Tamunodukobipi, A. Mamudu, and E. Goleiji, "Virtual safety engineer: from hazard identification to risk control in the age of AI," in *Safety-centric operations research: Innovations and integrative approaches: A multidisciplinary approach to managing risk in complex systems*: Springer, 2025, pp. 91-110.
[12] Q. Liu *et al.*, "AI-and Biotechnology-Driven Digital Design of Biohydrogen-Producing Microbiota," *Engineering,* 2025, doi: https://doi.org/10.1016/j.eng.2025.09.027.
[13] C. Gui, X. Zhao, S. Ai, X. Ouyang, A. Deng, and M. Li, "An augmented–depleted framework of AI awareness: Concept expansion and scale development," *International Journal of Hospitality Management,* vol. 128, p. 104164, 2025, doi: https://doi.org/10.1016/j.ijhm.2025.104164.
[14] L. C. Dias, A. Morton, and J. Quigley, "Elicitation: State of the art and science," *Elicitation: The science and art of structuring judgement,* pp. 1-14, 2017.
[15] M. Sethi and V. Verma, "Protecting Public Safety and Critical Infrastructure Systems in Smart Cities via Cybersecurity: AI-Based Threat Detection and Response," in *2025 Global Conference in Emerging Technology (GINOTECH)*, 2025: IEEE, pp. 1-6.
[16] D.-T. Vo, L. V. T. Nguyen, D. Dang-Pham, and A.-P. Hoang, "What makes an app authentic? Determining antecedents of perceived authenticity in an AI-powered service app," *Computers in Human Behavior,* vol. 163, p. 108495, 2025.
[17] R. Lakhdim, J. Treur, and P. H. Roelofsma, "Optimising blockchain security: Computational analysis of adaptive AI coaching," *Cognitive Systems Research,* p. 101430, 2025.
[18] A. Agarwal and M. J. Nene, "Addressing AI risks in critical infrastructure: formalising the AI incident reporting process," in *2024 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, 2024: IEEE, pp. 1-6.
[19] A. Yurdunkulu, M. A. Bulut, and A. Göçen, "From academics to Aidemics: Unpacking the human–AI symbiosis in higher education," *Acta Psychologica,* vol. 261, p. 105796, 2025.
[20] A. Lal and F. You, "Advances and challenges in energy and climate alignment of AI infrastructure expansion," *Advances in Applied Energy,* p. 100243, 2025.
[21] S. Hallaj *et al.*, "Open Data Sharing in Clinical Research and Participants Privacy: Challenges and Opportunities in the Era of Artificial Intelligence," *arXiv preprint arXiv:2508.01140,* 2025.
[22] I. J. Akpan, Y. M. Kobara, J. Owolabi, A. A. Akpan, and O. F. Offodile, "Conversational and generative artificial intelligence and human–chatbot interaction in education and research," *International Transactions in Operational Research,* vol. 32, no. 3, pp. 1251-1281, 2025.
[23] A. Setiyawan, S. Soeharto, T. T. Wijaya, L. Korenova, and Z. Lavicza, "Measuring Teachers' Competencies for AI Integration: Development and Validation of the AI-TPACK in Vocational Education," *Computers and Education Open,* p. 100319, 2025.
[24] M. Graves and E. Ratti, "A Capability Approach to Ethical Development and Internal Auditing of AI Technology," *Journal of Responsible Technology,* p. 100121, 2025, doi: https://doi.org/10.1016/j.jrt.2025.100121.

[25] V. Gallego, R. Naveiro, C. Roca, D. Rios Insua, and N. E. Campillo, "AI in drug development: a multidisciplinary perspective," *Molecular Diversity,* vol. 25, no. 3, pp. 1461-1479, 2021, doi: https://doi.org/10.1007/s11030-021-10266-8.
[26] S. Backman, "Risk vs. threat-based cybersecurity: the case of the EU," *European Security,* vol. 32, no. 1, pp. 85-103, 2023.
[27] A. Malik, M. T. De Silva, P. Budhwar, and N. Srikanth, "Elevating talents' experience through innovative artificial intelligence-mediated knowledge sharing: Evidence from an IT-multinational enterprise," *Journal of International Management,* vol. 27, no. 4, p. 100871, 2021, doi: https://doi.org/10.1016/j.intman.2021.100871.
[28] H. R. Hasan and K. Salah, "Combating deepfake videos using blockchain and smart contracts," *Ieee Access,* vol. 7, pp. 41596-41606, 2019, doi: DOI: 10.1109/ACCESS.2019.2905689.
[29] Z. Hu, M. Zhu, and P. Liu, "Adaptive cyber defense against multi-stage attacks using learning-based pomdp," *ACM Transactions on Privacy and Security (TOPS),* vol. 24, no. 1, pp. 1-25, 2020, doi: 10.1145/3418897.
[30] J. Camacho Sosa-Días, A. Couce-Vieira, D. Arroyo Guardeño, and D. Ríos Insua, "A cybersecurity risk analysis framework for systems with artificial intelligence components," 2025.
[31] Y. Gu, J. C. Goez, M. Guajardo, and S. W. Wallace, "Autonomous vessels: state of the art and potential opportunities in logistics," *International Transactions in Operational Research,* vol. 28, no. 4, pp. 1706-1739, 2021, doi: https://doi.org/10.1111/itor.12785.