electronics

Article

Standard-Oriented Architecture for AI-Powered Information Security Risk Management

Oleksii Chalyi [1,*], Kęstutis Driaunys [1], Šarūnas Grigaliūnas [1,2] and Rasa Brūzgienė [1]

[1] Kaunas Faculty, Institute of Social Sciences and Applied Informatics, Vilnius University, Muitines Str. 8, LT-44280 Kaunas, Lithuania; kestutis.driaunys@knf.vu.lt (K.D.); sarunas.grigaliunas@ktu.lt (Š.G.); rasa.bruzgiene@knf.vu.lt (R.B.)
[2] Department of Computer Sciences, Kaunas University of Technology, Studentu Str. 50, LT-51368 Kaunas, Lithuania
* Correspondence: oleksii.chalyi@knf.vu.lt

Abstract

This paper presents a standard-oriented architecture for automating information security risk management (ISRM) using artificial intelligence. The study first evaluates eight international frameworks (including COBIT 2019, NIST SP 800-53, and ISO 31000) for automation suitability, identifying ISO/IEC 27005 as the optimal structural foundation. Based on these findings, an architecture integrating Natural Language Processing and machine learning to automate risk identification, assessment, and treatment is proposed. A core component is a decision-making module that combines expert reasoning with a Multi-LLM consensus mechanism to ensure reliability. To provide exploratory support for the proposed architecture, a comparative study using five state-of-the-art Large Language Models (ChatGPT, Gemini Advanced, Grok, Microsoft Copilot, and DeepSeek Chat) was conducted on a standardized risk identification task. The results highlight strong cross-model consensus patterns, providing exploratory evidence that LLMs may support expert-informed risk identification and reasoning tasks while acknowledging the current limitations in complex reasoning. This approach proposes a transparent architectural foundation for AI-driven ISRM whose scalability must be established through future prototype-based evaluation, thereby bridging the gap between rigid compliance standards and generative AI capabilities.

Keywords: artificial intelligence; cybersecurity; information security; international standards; ISRM; risk analysis; risk management

Academic Editors: Yangjie Cao, Ziyang He and Minglin Liu

Received: 12 February 2026
Revised: 14 March 2026
Accepted: 16 March 2026
Published: 19 March 2026

1. Introduction

Due to the development of technologies and the implementation of everything into digital worlds, the question of information security (IS) has become very relevant and important [1]. In most cases, information security can be defined as the security of the CIA triad (confidentiality, integrity, and availability) [2]. However, information security is more about the status of information rather than directly securing it, which cybersecurity relates to. Cybersecurity involves processes and actions that can maintain information security [3]. As the number of cyber threats continues to grow, organizations must identify the most severe risks and apply effective countermeasures. Cybercrime is projected to cost $10.5 trillion in damages by 2025, a substantial leap from $3 trillion in 2015 [4]. Risk management provides a systematic approach to estimate the potential impact and likelihood of such threats, allowing risks to be classified as critical, high, medium, or low [5].

Prioritizing high risks and critical risks enables organizations to allocate resources efficiently. However, due to the vast number of potential threats, risk management remains a complex, time-consuming, and resource-intensive process that requires significant expertise in information security [6].

Artificial intelligence (AI), one of the most transformative technological advancements of the 21st century, enables machines to perform tasks traditionally requiring human intelligence [7–9]. In the context of cybersecurity, prior research has shown that AI can enhance information security risk management through advanced threat detection, predictive analytics, and automated response, though challenges remain around explainability, data dependency, and alignment with standards [10]. In addition to detection and response, AI has also been applied to risk prioritization in complex digital ecosystems, helping organizations allocate resources more effectively [11]. These applications highlight that the use of AI offers new opportunities to automate previously unmanageable processes, making cybersecurity risk management a particularly promising candidate for such automation.

The main objectives of this study are to analyze international standards for IS risk management (ISRM), evaluate the applicability of modern AI methods for automation, and propose a novel system architecture for AI-based ISRM.

Based on these objectives, the study addresses the following research questions (RQs):

- RQ1:Which information security risk management standard is best suited for automation using artificial intelligence?
- RQ2: How can ISO/IEC 27005 [12] risk management stages be mapped to specific AI methods and data sources?
- RQ3: What improvements in effectiveness, consistency, and coverage can be expected when applying the proposed AI-based architecture?

By addressing these research questions, this study makes a twofold contribution: it provides a comparative analysis of ISRM standards regarding their Automation Readiness and defines a reference architecture for implementing AI-driven risk management workflows.

2. Related Works

I. Hamid and M. M. H. Rahman (2025) provide a comprehensive analysis of the role of AI, machine learning (ML), and deep learning (DL) in modernizing cyber risk management [13]. Their study highlights how these technologies enhance vigilance by automating threat detection and reducing response times compared to traditional frameworks. Furthermore, the authors address critical regulatory and strategic implications, noting that while AI creates innovation opportunities, it introduces significant challenges regarding data privacy and model explainability. They emphasize that effective implementation requires a synergistic approach involving collaboration between AI experts, security professionals, and policymakers. However, while acknowledging these strategic needs, their work remains a high-level survey and does not propose a specific technical architecture for integrating these AI components into standard-compliant workflows like ISO/IEC 27005.

M. Yazdi et al. (2024), in their study, conducted a comprehensive analysis of the role of artificial intelligence in risk management [14]. They analyzed the ability of ChatGPT-4 to assess risks and compared these results with those of human experts. The comparison was based on six criteria. In the criteria of Completeness, Relevance, and Response Time, ChatGPT-4 scored higher compared to human experts. On the other hand, human experts received higher scores in the criteria of Practicality, Comprehensiveness, and Contextual Understanding. These results indicate that while ChatGPT-4 can provide fast answers, they are still not as complex and practical as those provided by human experts. In the conclusions, they highlighted several challenges, including accountability, data dependency,

inherent biases, the dynamic nature of risks, interpretability, and the difficulty of integrating AI systems.

N. Mohamed (2023) provides a quantitative analysis of the AI cybersecurity landscape, revealing a critical dichotomy: while 45% of organizations have already incorporated AI tools and 35% plan to do so, a significant segment (20%) remains hesitant [15]. The study identifies "decision-making transparency" and "algorithmic bias" as the primary barriers preventing full adoption. Mohamed explicitly calls for future research to focus on "strategies for responsible usage" and "unbiased prediction models" rather than just technical detection capabilities. However, while the paper accurately diagnoses these ethical and operational bottlenecks, it does not propose a specific governance framework to resolve them. This validates the necessity of the architecture proposed in our study, which addresses the transparency deficit not merely through post-factum documentation (Stage 4), but fundamentally by embedding Explainable AI (XAI) principles into its core decision-making engine. Specifically, the proposed system tackles algorithmic opacity through provenance tracking (logging structured decision rationales and evidence provenance, including explicit mappings between extracted threats and recommended controls), providing human-readable confidence scores for generated recommendations, and utilizing case-based reasoning to supply human-legible justifications derived from historical data. Coupled with bias mitigation via the Delphi consensus method, these mechanisms ensure that the AI acts as an interpretable, verifiable decision support tool.

S. S. Dasawat and S. Sharma (2023), in their paper, provide an overview of new-age sustainable startup businesses and the importance of emerging technologies in new-age sustainable startup businesses, risk management, automation, and scaling of entrepreneurs [16]. Furthermore, this paper discusses the impact of AI and cybersecurity on new-age sustainable startup businesses. The authors highlight that using AI can increase efficiency, provide better decision making, and predict maintenance needs. In conclusion, the authors state that the integration of artificial intelligence and automation can help businesses identify potential security threats and respond to them quickly and effectively. A combination of innovative technology, strategic planning, and effective risk management practices can help entrepreneurs effectively scale their businesses while also addressing the challenges of cybersecurity.

M. Sterbak et al. (2021), in their work "Automation of Risk Management Processes," provide a description of the individual subprocesses of information security risk management and identify the possibilities of applying automation to individual subprocesses and their interconnection within a complex information system [17]. Their proposed model is based on ISO/IEC 2700x standards and has the following main stages: the identification and description of the organization; the identification of information assets; the valuation of information assets; risk identification and assessment; mitigation; and monitoring and control. In conclusion, the authors state that from their analysis and the complexity of the problem, it is clear that it is not possible to fully automate all subprocesses and that some of them still require manual intervention. Among all stages, they mention that the automatic detection of IT assets with their subsequent categorization in a hierarchical model represents the greatest improvement and requirement. The current paper directly addresses this exact gap. While Sterbak et al.'s model relied on structured inputs and lacked the cognitive capabilities to process unstructured contextual data, the proposed architecture overcomes this limitation by integrating Natural Language Processing (NLP) and Large Language Model (LLM) reasoning. Specifically, instead of relying on manual intervention, proposed system utilizes LLM-based semantic parsing and Named Entity Recognition (NER) to automatically identify, extract, and categorize IT assets directly from unstructured organizational policies and architectural descriptions. This transition from rigid rule-based

models to dynamic AI reasoning provides the precise technological mechanism needed to fully automate the asset detection and categorization subprocesses that previous studies found unfeasible.

To provide a structured comparison of existing studies, Table 1 summarizes the main approaches, contributions, and limitations of the reviewed literature.

Table 1. Summary of related works.

Reference	AI Approach	Main Contribution	Limitations
I. Hamid & M. Rahman (2025) [13]	AI/ML/DL Survey	Defined strategic need for AI-Human synergy in cyber risk management	Remains a high-level survey; lacks a specific technical architecture for integrating AI into standard-compliant workflows
M. Yazdi et al. (2024) [14]	ChatGPT-4 vs. Human experts	Compared AI and human capabilities in risk assessment accuracy	AI outputs lacked the complexity and practicality of human experts; leaves challenges like accountability, data dependency, and inherent biases unaddressed.
N. Mohamed (2023) [15]	Quantitative Survey	Identified transparency and bias as key barriers (20% hesitation)	No governance framework proposed to address the transparency deficit; lacks specific strategies for responsible AI usage and unbiased prediction models.
S. S. Dasawat & S. Sharma (2023) [16]	Conceptual discussion	Highlighted AI's role in improving efficiency and early threat detection	Remains a conceptual discussion without technical implementation details or specific architectural guidelines.
M. Sterbak et al. (2021) [17]	ISO/IEC 2700x-based model	Identified automatable subprocesses in risk management	Concluded that full automation is unfeasible without AI reasoning, specifically highlighting the manual bottleneck of automatic IT asset detection and categorization.

All the authors emphasize the transformative impact of AI on risk management. Recent surveys confirm that while AI plays a vital role in modernizing security frameworks [13,16], its practical adoption faces hurdles. Specifically, relying solely on generative models like ChatGPT-4 lacks the contextual depth required for professional risk estimation [14]. Furthermore, while adoption rates are rising, significant hesitation remains due to the "transparency deficit" in AI decision making, as identified by quantitative studies [15]. Prior attempts to automate ISO/IEC 27005 processes [17] concluded that full automation is unfeasible without advanced reasoning capabilities. These findings highlight a critical gap: the lack of a unified, standard-oriented architecture that combines the reasoning power of LLMs with the auditability of formal risk management standards. This study aims to fill that gap.

3. Methodology

This study employs a comparative analytical approach to evaluate international standards relevant to information security risk management [18] and to identify the most suitable framework for automation using artificial intelligence. The research process consisted of two main phases: the comparative analysis of existing standards and the conceptual design of the proposed AI-based architecture.

To ensure a systematic and representative analysis, the following criteria were applied for the selection of standards:

- Inclusion Criteria: Globally recognized international standards or frameworks; current versions active as of 2025; those with a specific focus on risk management or information security governance; and those available in English.
- Exclusion Criteria: Deprecated standards (e.g., ISO/IEC 13335 [19]); frameworks strictly limited to specific national regulations without international applicability; and proprietary frameworks that are not publicly documented.

Based on this criteria, eight international standards and frameworks were selected: COBIT 2019 [20], ISO/IEC 27005:2023 [12], ISO 31000:2018 [21], The Standard of Good Practice [22], NIST SP 800-53 Rev.5 [23], IEC 62443-2-1 [24], COSO Internal Control—Integrated Framework [25], and the HITRUST Risk Management Framework [26]. The selection was guided by their frequency of citation in the academic literature, applicability to cybersecurity, and inclusion in national or organizational risk management policies. Each standard was evaluated using five criteria:

1. Structural Detail: The level of process decomposition, stage clarity, and procedural guidance. The more detailed the standard is (number of stages, sections, or procedural steps related to information security risk management), the higher score it receives.
2. ISRM Focus: The degree to which the standard supports ISRM processes. Standards addressing general risk management receive lower scores.
3. Automation Readiness: The extent to which the processes defined in the standard can be translated into algorithmic or data-driven workflows. Standards receive higher scores if they provide structured inputs (e.g., explicit asset values and threat frequencies), clear outputs (e.g., quantified risk matrices), and measurable parameters (e.g., distinct 1–5 probability scales). For instance, ISO/IEC 27005 maps these elements directly to mathematical variables, whereas frameworks like NIST SP 800-53 function more as qualitative control checklists, making direct algorithmic translation more complex.
4. Process Lifecycle Coverage: This measures whether the framework comprehensively covers all key stages of the ISRM lifecycle (e.g., context establishment, risk identification, analysis, evaluation, treatment, documentation, monitoring, and communication).
5. Compliance Assessment Support: This evaluates whether the framework provides clear metrics, checklists, or evaluation procedures that support automated compliance checking and decision making.

The qualitative assessment was conducted by a panel of four experts to minimize subjective bias. The panel composition included three academic researchers in information engineering and one certified information systems auditors. Each expert independently rated the standards on a five-point scale (1—minimal compliance; 5—full compliance) based on the rubric definitions.

To ensure inter-rater reliability, any score divergence greater than one point among the experts was analyzed and resolved through a guided consensus discussion. The evaluation utilized equal weighting for all five criteria. This equal weight approach was deliberately chosen because, in the context of developing a holistic AI-driven architecture, technical feasibility (Automation Readiness) is deemed equally as critical as domain relevance (ISRM Focus) and regulatory alignment (Compliance Assessment Support). The final score for each standard represents the averaged consensus of the expert evaluations. Furthermore, to verify the robustness of this selection, a conceptual sensitivity analysis was conducted post-scoring: because ISO/IEC 27005 received uniformly high scores (4 and 5) across all

dimensions, applying variable weights, such as increasing the weight of 'Automation Readiness' or decreasing 'Compliance Support', mathematically fails to displace it from the top rank. This consistency thoroughly validates ISO/IEC 27005 as the optimal objective foundation for the proposed architecture.

In the second phase, based on the findings of the comparative analysis, the ISO/IEC 27005 standard was selected as the foundation for developing an automated risk management architecture. The proposed architecture was modeled using the draw.io platform and designed to align with ISO/IEC 27005's four main stages: Context Establishment, Risk Assessment, Risk Treatment, and Documentation. Each stage was enhanced with AI-driven methods to demonstrate how automation can be applied within the standard's structure. Specifically, NLP is utilized in Context Establishment for the automated extraction of asset relationships and compliance requirements from organizational policies; ML is applied in Risk Assessment for predictive risk scoring and historical incident data analysis; and Multi-LLM decision support is integrated into Risk Treatment to evaluate and prioritize mitigation strategies. A comprehensive, granular breakdown of how these technologies map to each ISO/IEC 27005 activity, including data sources and expected outputs, is detailed in Section 4.2.

Additionally, to provide exploratory support for the proposed architecture, a comparative experiment was conducted using five state-of-the-art Large Language Model services on an ISO/IEC 27005-aligned risk identification task. The experiment utilized the premium consumer-facing web interfaces of these platforms accessed in November 2025: ChatGPT Plus (running the GPT-4o model, OpenAI), Gemini Advanced (running the Gemini 1.5 Pro model, Google), Grok (running the Grok-2 model, xAI), Microsoft Copilot (powered by GPT-4, Microsoft), and DeepSeek Chat (running the DeepSeek-V3 model, DeepSeek). It should be noted that because the evaluations were conducted via web interfaces to simulate a real-world, off-the-shelf usage scenario by a risk analyst, the exact backend API model versions and iterative weights are opaque and managed dynamically by the vendors. Therefore, the results reflect the capabilities of the flagship models deployed by these services as of November 2025. These specific models were deliberately selected to provide a comprehensive cross-section of the current state-of-the-art landscape. They represent a diverse mix of architectural approaches and access paradigms. By including models from different corporate ecosystems and training paradigms, the experiment ensures that any observed consensus in risk reasoning is robust and not merely an artifact of a single vendor's specific alignment or architectural bias.

Each model was tasked to analyze the open-access security policy template provided by the Center for Internet Security (CIS) [27]. To evaluate the models' intrinsic reasoning capabilities without guided examples, a "Zero-Shot Role Prompting" strategy was employed. This approach simulates a real-world scenario where an organization lacks pre-labeled training data. The unified prompt used for all models was as follows:

> *"Based on the following company security policy, identify the ten most potential information security risks according to the ISO/IEC 27005 classification. Provide each risk with a short description and its category (e.g., confidentiality, integrity, availability). The result should be a table with four columns: Risk number, Risk name, Risk influence on CIA, and Risk description."*

The number of identified risks was limited to ten to ensure comparability and to focus on the models' ability to prioritize and reason about the most critical threats.

Rather than benchmarking against a predefined "correct" list, the experiment aimed to reveal differences in reasoning patterns, prioritization, and CIA-triad interpretation across LLMs. These variations were analyzed to understand how different models conceptualize risk within the ISO/IEC 27005 framework.

This analysis directly supports the design of the proposed AI-driven Decision-Making System, which combines multiple LLMs and expert reasoning in a Delphi-style aggregation. Understanding inter-model divergence helps to identify complementary reasoning strengths and bias patterns, improving the ensemble's ability to reach balanced and explainable risk assessments. Consequently, the experiment provides exploratory evidence regarding the potential use of LLMs for ISO/IEC 27005-aligned risk identification and informs their proposed role as reasoning agents within the architecture's decision support layer, without constituting a full empirical validation.

4. Results

This section presents the main outcomes of the study, beginning with a comparative analysis of international information security risk management standards, followed by the development of a proposed architecture derived from these findings. Additionally, an experimental evaluation using five state-of-the-art Large Language Models was conducted to assess their capability to identify and classify information security risks according to ISO/IEC 27005 principles.

4.1. Information Security Risk Management International Standards Comparison

One of the international standards that describe risk management is the COBIT 2019 Framework [20]. COBIT is a comprehensive framework for the governance and management of information and technology, designed to support enterprises as a whole. The framework is structured around governance and management objectives. One of the four domains of management objectives is Align, Plan, and Organize (APO), which addresses the overall organization, strategy, and supporting activities for I&T. Within this domain, APO12 Managed Risk outlines the key aspects of risk management. This framework divides the management practice into six key activities:

1. APO12.01 Collect Data: This describes where to find and which data to collect in order to analyze the risks. It recommends capturing relevant data from related issues, incidents, problems, and investigations.
2. APO12.02 Analyze Risk: This describes the important steps of risk estimation, such as identifying the list of risks, estimating their frequency (or probability), and finding the magnitude of loss. It also proposes analyzing the cost or benefit of potential risk response options, such as avoiding, reducing, transferring, and accepting and exploiting/seizing.
3. APO12.03 Maintain a Risk Profile: This involves maintaining an inventory of known risks and risk attributes, including expected frequency, potential impact, and responses.
4. APO12.04 Articulate Risk: This involves communicating information on the current state of I&T-related exposures and opportunities in a timely manner to all required stakeholders for an appropriate response.
5. APO12.05 Define a Risk Management Action Portfolio: This involves defining a balanced set of project proposals designed to reduce risk.
6. APO12.06 Respond to Risk: This describes the response plan to minimize the impact when risk incidents occur.

The COBIT Framework provides valuable guidance for cybersecurity risk management, offering a general understanding of risks' lifecycles and emphasizing governance integration within enterprise IT environments.

Another key international framework addressing risk management is ISO/IEC 27005, "Information technology–Security techniques–Information security risk management" [12]. This standard is part of the ISO/IEC 27000 series and is completely dedicated to risk man-

agement in information security. While ISO/IEC 27001 [28] also includes risk management, it does so at a general level, whereas ISO/IEC 27005 provides an in-depth, comprehensive approach that covers all stages of the process. The standard has undergone four major versions: 2008, 2011, 2018, and 2022. In the latest 2022 version, the structure of the information security risk management process was updated. According to this version, the process consists of four core stages:

1. Context Establishment, which involves defining criteria for risk evaluation, impact, and acceptance; determining scope and boundaries; and organizing the overall risk management process.
2. Risk Assessment, which consists of risk identification (including assets, threats, controls, vulnerabilities [29], and potential consequences), risk analysis (using qualitative and/or quantitative assessment methods to estimate probabilities and impacts), and risk evaluation.
3. Risk Treatment, which involves selecting appropriate actions to reduce, avoid, accept, or transfer risks.
4. Documented Information, which involves replacing the former "Risk Acceptance" stage from previous versions. It emphasizes recording and reporting processes to ensure proper documentation and traceability.

Additionally, two continuous activities—Monitoring and Review, and Communication and Consultation—are performed throughout the first three stages of the information security risk management process. Unlike the COBIT Framework, ISO/IEC 27005 provides explicit recommendations and detailed procedural guidance for each phase of risk management. It serves as a step-by-step reference for cybersecurity risk managers, offering structured and actionable instructions for implementing an effective risk management process.

ISO/IEC 31000:2018 "Risk management: Guidelines" is another international standard that describes risk management [21]. Unlike ISO/IEC 27005, ISO/IEC 31000:2018 is more similar to COBIT 5 as it provides a general description and does not offer specific and detailed information. However, the relationships between risk management principles, frameworks, and processes outlined in ISO/IEC 31000:2018 can be useful for risk managers. The main processes of risk management according to this standard are similar to those described in ISO/IEC 27005 and consist of the following stages:

1. Scope, Context and Criteria: This involves defining the scope of the process, and understanding the external and internal context.
2. Risk Assessment: This encompasses risk identification, risk analysis and risk evaluation.
3. Risk Treatment: This involves selecting and implementing appropriate options for addressing risks.
4. Recording and Reporting: This involves documenting the risk management process and its outcomes through appropriate mechanisms.
5. Communication and Consultation: This relates to assisting relevant stakeholders in understanding risk, the basis on which decisions are made and the reasons why particular actions are required. This process should take place within and throughout all steps of the risk management process.
6. Monitoring and Review: This involves assuring and improving the quality and effectiveness of the process's design, implementation and outcomes. This process should take place at all stages of the process.

This standard focuses on general risk management principles and does not specifically address information security risk management.

One more international framework related to information security risk management is The Standard of Good Practice (SOGP) for Information Security [22]. The standard

includes four sections that are directly related to information risk management: two from the Security Management (SM) domain, one from Critical Business Applications (CB), and one from Computer Installations (CI).

1. SM3.3 Managing information risk analysis: This describes how to identify key information risks and determine the controls required to keep those risks within acceptable limits.
2. SM3.4 Information risk analysis methodologies: This provides a general description of information risk analysis methodologies and describes its application.
3. CB5.3 Information risk analysis: This identifies key information risks associated with the application, and defines the security controls required in order to keep those risks within acceptable limits.
4. CI5.4 Information risk analysis: This identifies key information risks associated with the computer installation and defines the security controls required in order to keep those risks within acceptable limits.

The SOGP provides concise theoretical guidance on information risk management and can be useful for strengthening the documentation and analytical stages of the risk management process.

The National Institute of Standards and Technology (NIST), in its Special Publication 800-53 Rev. 5: Security and Privacy Controls for Information Systems and Organizations [23], provides comprehensive guidance for cybersecurity risk management, similar in scope to ISO/IEC 27005. This standard defines a structured list of control families designed to manage security and privacy controls across information systems and organizations. One of these control families, the Risk Assessment (RA) Family, includes nine key control elements:

1. Policy and Procedures: Establish security and privacy assurance foundations.
2. Security Categorization: Describes the potential adverse impacts on organizational operations, assets, and individuals.
3. Risk Assessment: Evaluates threats, vulnerabilities, likelihood, and impact, including risks from external parties.
4. Vulnerability Monitoring and Scanning: Ensure that potential sources of vulnerabilities, such as infrastructure components, networked devices, and peripherals, are not overlooked.
5. Technical Surveillance Countermeasures Survey: Detects the presence of technical surveillance devices or hazards and identifies related security weaknesses.
6. Risk Response: Determines appropriate responses to identified risks before generating action plans or milestones.
7. Privacy Impact Assessments: Evaluate privacy risks associated with information systems or activities and propose mitigation measures.
8. Criticality Analysis: Supports the prioritization of protection activities.
9. Threat Hunting: Proactively searches organizational systems, networks, and infrastructure for advanced or persistent threats.

The NIST SP 800-53 publication provides detailed, actionable guidance for implementing cybersecurity risk management processes and controls. However, compared to ISO/IEC 27005, it offers fewer practical examples and less emphasis on the procedural flow of the overall risk management lifecycle.

The International Electrotechnical Commission (IEC) 62443 Part 2-1: "Establishing an Industrial Automation and Control System Security Program" [24] is an international standard that describes risk management within the context of a Cybersecurity Management System (CSMS). This standard divides risk management into three key categories:

1. Risk Analysis: This discusses the background information that supports other CSMS components, divided into two elements: business rationale, and risk identification, classification, and assessment.
2. Addressing Risk with the CSMS: This contains the main requirements and information forming the core of the CSMS, structured into three groups: security policy, organization, and awareness; selected security countermeasures; and implementation.
3. Monitoring and Improving the CSMS: This focuses on ensuring that the CSMS is properly applied and continuously enhanced. It consists of two elements: conformance and the review, improvement, and maintenance of the CSMS.

IEC 62443 Part 2-1 provides detailed guidance on managing cybersecurity risks within industrial environments and explains the structure and lifecycle of a Cybersecurity Management System in depth.

Another globally recognized framework is the Internal Control—Integrated Framework, developed by the Committee of Sponsoring Organizations of the Treadway Commission (COSO IC-IF) [25]. Unlike other standards, it provides very detailed information regarding risk assessment rather than risk management. This framework highlights four main principles related to the risk assessment component:

1. The organization specifies objectives with sufficient clarity to enable the identification and assessment of risks relating to objectives.
2. The organization identifies risks to the achievement of its objectives across the entity and analyzes risks as a basis for determining how the risks should be managed.
3. The organization considers the potential for fraud in assessing risks to the achievement of objectives.
4. The organization identifies and assesses changes that could significantly impact the system of internal control.

This framework states that risk assessment is a dynamic, ongoing process that identifies and evaluates internal and external threats to organizational objectives against established risk tolerances. It depends on clearly defined goals across operations, reporting, and compliance, and requires management to account for environmental and business model changes that could undermine internal control. Compared to ISO/IEC 27005 or NIST SP 800-53, the COSO IC-IF framework focuses more on enterprise-level control and organizational governance rather than technical implementation. However, its structured approach to identifying and assessing risks provides valuable conceptual support for integrating risk management automation into broader corporate control systems.

The HITRUST Risk Management Framework (RMF) [26] was originally developed as a best-practice approach for the healthcare industry but has since evolved into a widely recognized cross-industry standard. In this study, it was also analyzed to identify its approach to risk management and its key processes. The framework defines risk management as a program and the supporting processes designed to manage information security risks to organizational operations, assets, individuals, and other entities. It includes four key components: (i) establishing the context for risk-related activities; (ii) assessing risk; (iii) responding to identified risks; and (iv) continuously monitoring risk over time. HITRUST RMF proposes a four-step model supported by guiding questions for each stage:

1. Identify risks and define requirements:“What are my protection requirements?”
2. Specify controls:“How do I provide the protection?”
3. Implement and manage controls:“Provide the protection.”
4. Assess and report:“How is my protection working?”

This structured, question-driven approach ensures a continuous and measurable risk management cycle. While HITRUST RMF is compliance-oriented and primarily used in

regulated sectors such as healthcare, its integration of multiple international standards (including ISO/IEC 27001, NIST SP 800-53, and COBIT) makes it a valuable framework for organizations seeking unified and auditable risk governance.

Eight international standards were analyzed that can be applied to risk management, including cybersecurity risk management for some. To complement this technical comparison, Figure 1 illustrates the relative popularity and the visibility of the analyzed standards based on their presence in the academic literature and web search results. This provides an additional perspective on the global adoption and recognition of each framework.

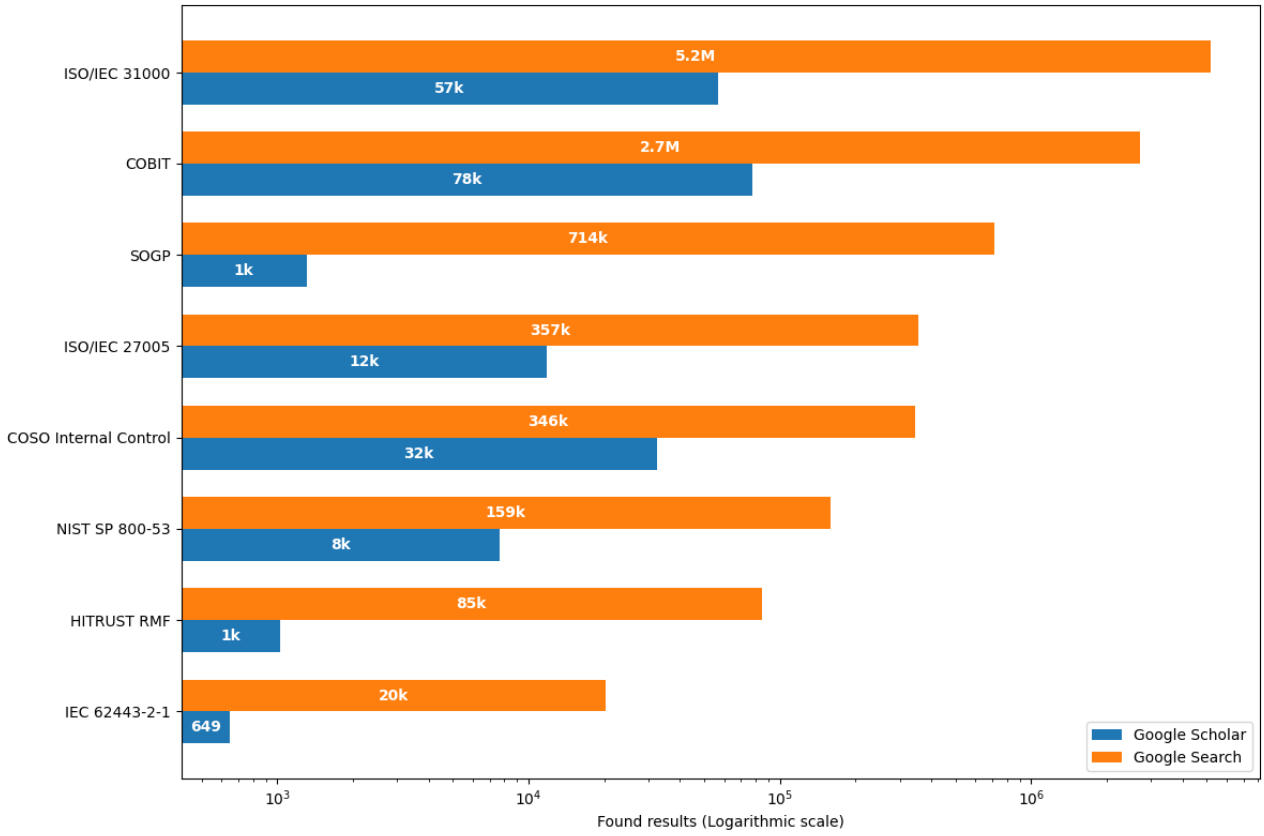

Figure 1. Popularity of risk management frameworks based on Google Scholar and Search results.

The detailed comparative evaluation of the standards across all the defined criteria is summarized in Table 2.

Table 2. Comparative evaluation of ISRM standards by key criteria.

Framework	Structural Detail	ISRM Focus	Automation Readiness	Process Lifecycle Coverage	Compliance Assessment Support	Final Score
COBIT 2019 [20]	3	3	2	5	5	3.60
ISO/IEC 27005 [12]	5	5	4	5	4	4.60
ISO 31000 [21]	3	3	4	5	5	4.00
SOGP [22]	1	3	2	3	2	2.20
NIST SP 800-53 [23]	4	4	3	5	5	4.20
IEC 62443-2-1 [24]	3	4	3	4	3	3.40
COSO IC-IF [25]	4	3	2	4	3	3.20
HITRUST CSF [26]	4	3	3	4	3	3.40

Among all frameworks, ISO/IEC 27005 achieved the highest overall score (4.60), confirming its dominance as the most suitable foundation for this study. It received maximum scores (5) in Structural Detail, ISRM Focus, and Process Lifecycle Coverage, reflecting its comprehensive nature as a dedicated risk management standard. Furthermore, it demonstrated the highest Automation Readiness (4), making it the most technically actionable framework for algorithm-driven implementation. Although its Compliance Assessment Support (4) is slightly lower than strictly control-based standards, its procedural clarity outweighs this factor for the purpose of architectural design.

The NIST SP 800-53 framework ranked second with a total score of 4.20. It demonstrated exceptional strength in Process Lifecycle Coverage (5) and Compliance Assessment Support (5), driven by its extensive control catalog. However, its lower score in Automation Readiness (3) compared to ISO/IEC 27005 indicates that while it is ideal for checking compliance, it is less adapted for automating the generative reasoning required in risk analysis.

ISO 31000 followed with a score of 4.00. While it achieved perfect scores in Process Lifecycle Coverage (5) and Compliance Assessment Support (5), its generalist nature resulted in lower scores for Structural Detail (3) and ISRM Focus (3). It lacks the specific technical granularity required to ground a cybersecurity-specific AI model.

COBIT 2019 achieved a score of 3.60. Its strong governance focus yielded maximum scores in Process Lifecycle Coverage (5) and Compliance Assessment Support (5). However, its low Automation Readiness (2) and moderate Structural Detail (3) limit its utility as a core engineering specification for the proposed system.

Both IEC 62443-2-1 and HITRUST CSF achieved an identical score of 3.40. IEC 62443-2-1 showed balanced performance with a strong ISRM Focus (4) but moderate scores elsewhere, reflecting its niche applicability in industrial environments. HITRUST CSF demonstrated high Structural Detail (4) but average Automation Readiness (3), constrained by its heavy compliance orientation which is less flexible for dynamic AI reasoning.

The COSO IC-IF scored 3.20, showing strong Structural Detail (4) but weak Automation Readiness (2). As an enterprise-level internal control framework, it lacks the specific technical workflows necessary for automated information security risk treatment.

Finally, the SOGP showed the lowest total score (2.20), suffering from minimal Structural Detail (1) and low Compliance Assessment Support (2). Its methodology, while functional, does not provide the depth required for a modern, automated architecture.

In summary, the comparative analysis demonstrates that ISO/IEC 27005 outperforms all other standards in methodological completeness and technical adaptability. While NIST SP 800-53 and ISO 31000 offer powerful compliance and lifecycle coverage, ISO/IEC 27005 provides the optimal balance of structure (5) and Automation Readiness (4). This makes it the most appropriate baseline for developing the AI-driven ISRM architecture proposed in the following section.

4.2. Proposed Architecture

Based on the findings from the previous section, where ISO/IEC 27005 was identified as the most suitable standard for automation using AI techniques, the proposed architecture illustrated in Figure 2 was developed.

The proposed architecture aligns with the four core stages of information security risk management defined in ISO/IEC 27005, which are referenced as steps (1–4) in Figure 2.

Stage 1: Context Establishment. At this stage, the user defines the criteria for risk evaluation, impact, and acceptance, along with the scope and boundaries of the process. This requires interaction with internal sources such as organizational security policies and procedural documentation. To automate this step, Natural Language Processing methods are proposed to analyze and extract relevant information from internal documents.

Stage 2: Risk Assessment. This stage includes three key subprocesses: risk identification, risk analysis, and risk evaluation.

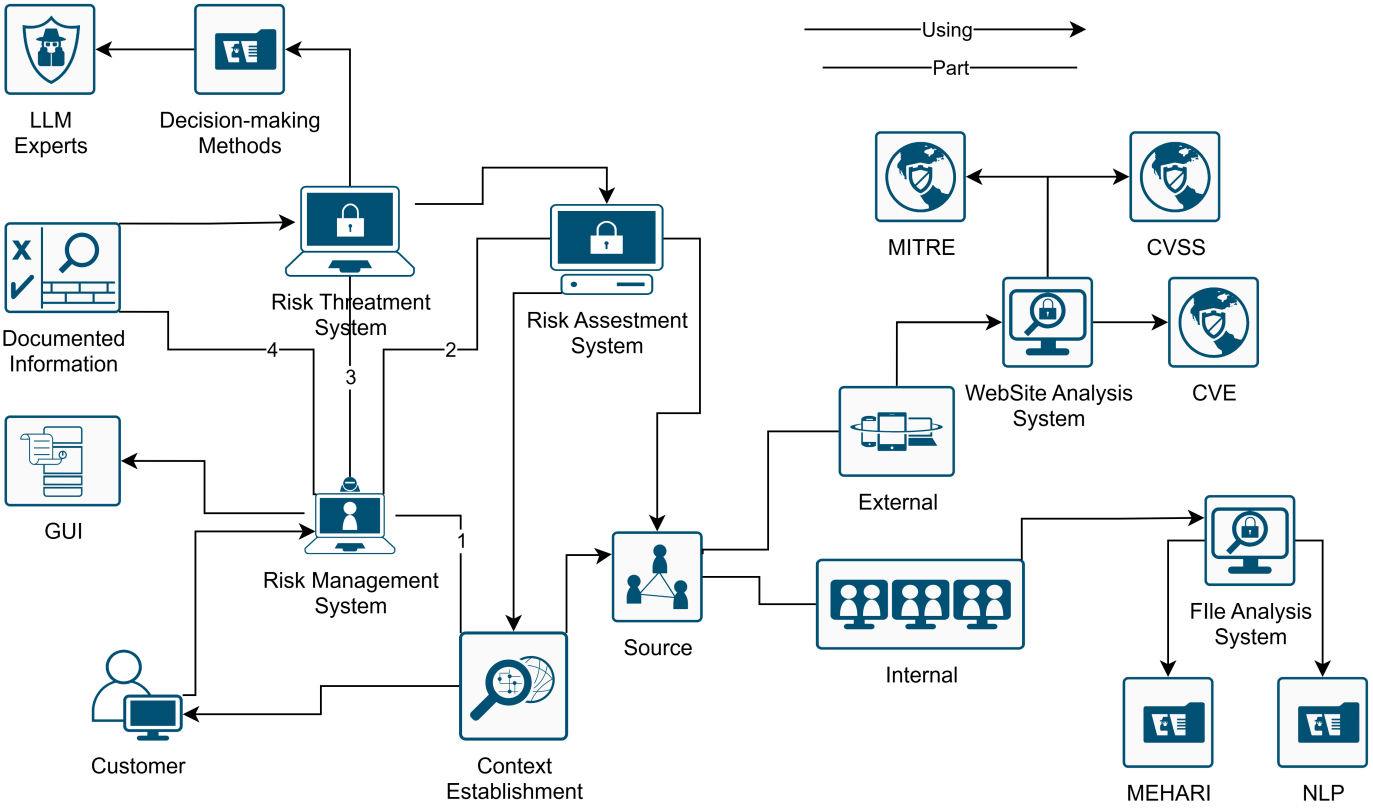

Figure 2. Proposed architecture of the AI-powered information security risk management system.

For risk identification, NLP techniques will be employed to analyze existing company risk reports, if available. In cases where such documentation is absent, the MEHARI (MEthod for Harmonized Analysis of Risk) project [30] can serve as a methodological reference. Developed by CLUSIF and CLUSIQ in 2010, MEHARI represented one of the earliest practical attempts to automate risk management using Excel-based tools derived from ISO/IEC 27005 and 27002 [12,31]. Although no longer updated since 2013, it demonstrated the feasibility of practical ISRM automation and remains valuable for conceptual grounding.

For risk analysis (severity to risk inputs), vulnerability severity is distinguished from organizational risk. Where a CVE exists, CVSS components are extracted as signals, not as a risk score [32]. The Exploitability-related factors (e.g., attack vector/complexity, required privileges, and user interaction) inform the Likelihood term; the technical impact factors (confidentiality, integrity, and availability) inform the Impact term after weighting by asset criticality and CIA requirements defined in Stage 1. Where no CVE exists, equivalent signals are derived from MITRE ATT&CK techniques and procedures, internal event frequency and exposure data, and residual control effectiveness. NLP is used to normalize unstructured evidence into these factors.

For risk evaluation with documented scales each risk scenario $s = \langle$asset, threat event, vulnerability, and controls$\rangle$ R_s is computed as $L_s \times I_s$, with scales and thresholds declared in Context Establishment (risk criteria, CIA weights, and acceptance thresholds). Likelihood (L_s) combines threat event frequency (internal incidents/threat intel), exploitability (CVSS derived where available, otherwise ATT&CK derived), exposure (attack surface; time to patch), and residual control effectiveness. Impact (I_s) combines asset criticality with CIA

effects using the organization's CIA weights. Both L and I are calibrated on 1–5 ordinal scales. R_s is categorized (Low/Med/High/Critical) per acceptance criteria set in Stage 1.

Stage 3: Risk Treatment. Treatment selection is cast as a budget- and policy-constrained optimizationproblem, supported by case-based reasoning (CBR), with an optional, offline reinforcement learning (RL) extension for control sequencing. RL is not required for the architecture to operate and is included only as an optional enhancement when sufficient logged data or a validated simulator is available. This keeps the decision engine auditable (optimization/CBR) and uses learning only where sufficient logged data or safe simulation exists (compare with the "Decision-making Methods" and "Risk Treatment System" blocks in Figure 2.

RL setup, simulator, and off policy evaluation.

1. Environment and reward. The RL environment uses the state/action/reward defined in Stage 3. Risk R is computed as in Stage 2 (Likelihood $\times$ Impact with Stage 1 criteria). Rewards reflect organizational objectives via λ weights published in the risk criteria.
2. Data sources (ground truth). Logged tuples$(s_t, a_t,$ and outcomes$_{t+1})$ come from change records, control deployments, red team exercises, vulnerability closure data (exposure/time to patch), and incident tickets. Where logs are insufficient, a data-driven simulator calibrated to historical frequencies and control effectiveness yields counterfactual outcomes for safe training.
3. Offline learning and safety. Policies are trained with offline RL/bandits using logged data. Before any deployment, an off-policy evaluation and A/B sandboxes are run. Actions are restricted to compliance safe sets, and budget and capacity constraints are enforced at selection time.
4. Baselines. The RL policy is benchmarked against constrained optimization (primary) and CBR retrieval + adaptation, with NDCG on prioritized treatments, ΔRisk, and the time to threshold reported.

Primary Method: Constrained Optimization. Let $x_j \in \{0,1\}$ indicate selection of ISO/IEC 27002 control j, let ΔR_s be the estimated reduction in organizational risk for scenario s (from Stage 2), and let $E_{js} \in [0,1]$ denote the effectiveness of control j on scenario s (derived from expert mappings, ATT&CK relations, and past outcomes). The controls are selected by solving

$$\max_{x} \quad \sum_{s} \sum_{j} E_{js}\, \Delta R_s\, x_j \tag{1}$$

$$\text{s.t.} \quad \sum_{j} \text{Cost}_j\, x_j \leq B, \tag{2}$$

$$\sum_{j} \text{Effort}_j\, x_j \leq H, \tag{3}$$

$$x \in \mathcal{C}, \tag{4}$$

where B is the budget, H the implementation capacity, and $\mathcal{C}$ encodes compliance constraints and risk acceptance thresholds declared in Stage 1 (e.g., must-have controls, segregation-of-duties rules, and target residual risk categories). The solver yields a prioritized, auditable treatment plan with projected risk reduction per unit cost/effort.

Support Method: Case-Based Reasoning. A retrieval and adaptation step surfaces past cases ⟨context, risk portfolio, controls, and outcomes⟩ similar to the current situation. Their observed outcomes seed E_{js}, inform constraint $\mathcal{C}$ (e.g., change windows and dependencies), and provide human-legible justification for recommended actions.

Optional Extension: Offline RL for sequencing (well scoped). Additionally, an offline RL policy is evaluated that learns control sequencing under the same constraints using logged decisions and a validated simulator.

- State s_t: Snapshot of residual risks (Stage 2), asset criticality and CIA weights (Stage 1), implemented controls, exposure, and remaining budget/capacity.
- Action a_t: Choose a treatment option (avoid/reduce/transfer/accept); for "reduce", select control j (and parameters, if applicable).
- Transition: Estimated from post-change outcomes and a data-driven simulator that models incident rates and residual risk given control sets and exposure.
- Reward (organizational, not CVSS):

$$\begin{aligned} r_t \;=\; & \lambda_1 \left(R_t^{\text{pre}} - R_{t+1}^{\text{post}}\right) \;-\; \lambda_2 \,\text{Cost}(a_t) \\ & -\; \lambda_3 \,\text{OpsImpact}(a_t) \\ & -\; \lambda_4 \,\text{NonCompliancePenalty}_{t+1}, \end{aligned} \tag{5}$$

where R is the Stage 2 organizational risk (Likelihood $\times$ Impact with Stage 1 criteria), and λ are policy weights fixed during Context Establishment. A small action change penalty can be added to discourage thrashing; discounting favors earlier risk reduction.

Ground truth and evaluation. Ground truth for learning and calibration comes from (i) post-implementation outcomes (incident frequency/severity, MTTD/MTTR, loss proxies/ALE, and residual risk trajectories) and (ii) a validated simulator when historical coverage is sparse. Minimum data quality assumptions are required before enabling the offline RL module. Logged data should be sufficiently complete to reconstruct state–action–outcome tuples, temporally consistent to preserve the ordering between exposure, control deployment, and observed outcomes, and explicitly linked across assets, controls, and incident records. In addition, incident labels and closure outcomes should be sufficiently consistent to support reliable reward estimation and case comparison. When historical coverage is sparse, the simulator is calibrated to historical incident frequency, estimated control effectiveness, and exposure windows derived from patch latency and control activation timing. If these minimum data quality or calibration conditions are not met, the RL module is disabled and the architecture falls back to the auditable constrained optimization and case-based reasoning pathway only. Before any online use, policies are compared via off-policy evaluation on logged data. The key performance indicators (KPIs) include ΔRisk vs. baseline, the time to acceptable risk, budget adherence, implementation success rate, and policy value estimates. Table 3 summarizes inputs, methods, outputs, and KPIs for this stage.

Stage 4: Documented Information. In this stage, the system automatically generates a report consolidating results from previous stages. The report will include the following:

- A risk evaluation table from Stage 2.
- Prioritized recommendations for risk treatment from Stage 3.

This report ensures traceability and compliance with ISO/IEC 27005 documentation requirements.

For the validation of results, experts in information security will be involved, and the Delphi method will be applied to obtain independent and anonymous evaluations from cybersecurity specialists. This will ensure the verification of the practical feasibility of the proposed approaches and the comparison with existing methods. In addition, experimental testing of the effectiveness of these approaches in a test environment is planned. To determine the system's effectiveness, developed criteria such as accuracy and speed of risk classification will be applied. To obtain the weights of these criteria, the Saaty

method or the pairwise comparison method will be used. To illustrate how ISO/IEC 27005 stages were operationalized within the proposed AI-powered framework, Table 3 presents the mapping between standard-defined activities, input data sources, applied AI methods, expected outputs, and evaluation metrics.

Table 3. Standards-to-AI mapping between ISO/IEC 27005 activities and proposed AI components.

ISO/IEC 27005 Activity	Data/Sources Used	AI Method/Technology	Output	Proposed Evaluation Metric (KPI)
Stage 1: Context Establishment	Internal policies, regulatory documents, and organizational IS descriptions	NLP (text classification and semantic parsing)	Identified risk domains, operational context, and asset inventory	Coverage accuracy (%)
Stage 2a: Risk Identification	CVE database, MITRE ATT&CK, incident logs, and SOC reports	Named Entity Recognition; LLM-based threat extraction	Structured list of potential threats and vulnerabilities	Risk coverage level (%)
Stage 2b: Risk Analysis	Asset values, historical incident data, threat likelihoods, and impact scales	Machine Learning regression, Bayesian classification, and Big Data Analytics	Quantified risk scores and probability impact matrices	RMSE; mean absolute error
Stage 2c: Risk Evaluation	Historical risk records; expert feedback datasets	LLM-based evaluation model + Delphi method	Prioritized and validated risk list	Expert agreement coefficient (κ); consistency ratio
Stage 3: Risk Treatment	ISO/IEC 27002 control catalog, risk portfolio, incident/outcome logs, and cost data	Constrained optimization (ILP), Case-Based Reasoning, and Offline Reinforcement Learning	Prioritized control set, rationale from similar cases, and projected risk reduction	ΔRisk vs. baseline, time to acceptable risk, and budget adherence
Stage 4: Documented Information	Consolidated system outputs from previous stages	NLP-based report synthesis and summarization	Automatically generated ISRM report	Completeness ratio; user satisfaction (Likert scale)

4.3. Illustrative Walkthrough: Backup Policy Gap Analysis

To demonstrate the practical workflow, a scenario where the system analyzes an organization's "Information Security Policy" is considered. The process follows the four proposed stages:

Stage 1 (Context Establishment): The NLP module parses the policy text and detects a compliance gap in the Section 4.4. While the policy states "Backups must be performed," it fails to define the Frequency (RPO) and Off-site Storage requirements. The system flags this ambiguity as a vulnerability against ISO/IEC 27001 control A.12.3.1.

Stage 2 (Risk Assessment): Instead of relying on generative estimation, the system calculates the organizational risk score using a deterministic likelihood impact model defined in Stage 1. Mapping the "Data Loss" threat to the critical asset "Customer Database" (Availability Requirement: High), CVSS exploitability and technical impact vectors are used as structured input signals rather than as the final risk score. These signals inform the Likelihood term, while asset criticality and CIA requirements inform the Impact term. The final organizational risk score is then computed according to the Stage 1 criteria and identifies this scenario as a high-priority risk requiring immediate attention.

Stage 3 (Risk Treatment): This stage utilizes the Multi-LLM Decision-Making Module to generate and select the optimal mitigation strategy. Five independent LLM agents propose treatment plans; for instance, Model A suggests "Real-time mirroring" (High Cost), Model B suggests "Daily incremental backups to Cloud" (Medium Cost), while Model C

proposes "Weekly local backups" (Low Effectiveness). Using the Delphi method, the system aggregates these proposals. Through iterative rounds, the ensemble rejects Model C (non-compliant with off-site needs) and Model A (violates the budget constraint). The consensus converges on Model B's approach, and the mathematical solver (Equations (2)–(5)) confirms that "Daily Cloud Backups" maximizes risk reduction within the available resources.

Stage 4 (Documented Information): The system automatically generates a "Risk Treatment Plan," detailing the selected backup schedule and storage provider, ready for CISO approval.

4.4. Operational Governance and AI Safety Mechanisms

To address the challenges of accountability and reliability in AI-driven workflows, the architecture integrates specific governance mechanisms aligned with the ISO/IEC 27005 process flow:

(1) Human-in-the-Loop Validation:While the system automates data processing and initial reasoning, the final decision authority remains with human experts, particularly during the Risk Treatment stage. High-impact decisions (e.g., accepting a critical risk or allocating significant budget for controls) trigger a mandatory manual review workflow. The AI acts as a decision support agent, providing recommendations with confidence scores, rather than an autonomous actor. Consequently, the operational and legal accountability for any implemented risk treatment, including those potentially resulting in financial or regulatory impacts, rests entirely with the human decision maker (e.g., the CISO or Risk Manager), with the AI serving strictly as an auditable advisory tool.

(2) Anti-Hallucination Strategy: To minimize the risk of generative fabrication, the LLM's reasoning is strictly bounded by the definitions set in Stage 1: Context Establishment. The model is configured to treat the ingested organizational policies and the external threat databases (MITRE; CVE) as the sole source of truth, rejecting any inference that cannot be traced back to these inputs. This ensures that identified risks are contextually valid rather than hallucinated.

(3) Auditability, Explainable AI and Data Governance: The Documented Information stage serves as a comprehensive audit trail integrated with XAI principles. Beyond generating the final risk report, this component ensures strict provenance tracking by recording structured decision rationales and evidence provenance—specifically, the exact mapping between the detected threat (Stage 2), the supporting sources, and the selected control (Stage 3), together with model confidence scores. This transparency allows human auditors to verify whether a decision was derived from algorithmic logic, historical case-based reasoning, or expert input, directly satisfying the requirement for interpretable AI.

(4) Data Privacy and Deployment Security: Given the sensitivity of internal security policies and vulnerability data, the architecture mandates a strictly isolated deployment model. For highly sensitive environments, the system utilizes open-weight LLMs (e.g., Llama 3 and Mistral) deployed within the organization's secure perimeter, ensuring no data leaves the internal network.

(5) Ethical Compliance and Bias Mitigation: The use of AI in risk management introduces potential biases (e.g., models over-emphasizing technical threats while underestimating social engineering due to training data skew). To address this risk, the Decision-Making System utilizes an ensemble approach (Delphi method) across multiple diverse models to reduce the dependence on any single model's idiosyncratic reasoning and to broaden the range of perspectives considered during risk evaluation.

In this context, "diverse models" refers to employing LLMs developed by different vendors (e.g., OpenAI's GPT, Google's Gemini, etc.). Because these models utilize different pre-training corpora, architectural configurations, and alignment strategies, they may

expose different blind spots and reasoning patterns. However, shared or overlapping biases may still persist, especially where training data sources or alignment tendencies are similar. The technical implementation of the AI Delphi method operates as an iterative, multi-agent evaluation mechanism:

Phase 1 (Independent Generation): Each model is queried independently with a prompt to generate its risk assessment or treatment plan without influence from the others.

Phase 2 (Cross-Evaluation): The outputs are aggregated, anonymized, and fed back to the models. Each model is prompted to review the peer assessments, identify potential blind spots or over-exaggerated risks in its own initial output, and refine its response based on the collective reasoning.

Phase 3 (Consensus Aggregation): A final consensus is derived mathematically. A risk or treatment is accepted only if it meets a predefined consensus threshold (e.g., four/five models agree on the risk classification). Disputed risks are flagged for human review.

To ensure verifiability, this entire iterative process, including initial divergence, peer critique, and final consensus scoring, is logged within the Documented Information module (Stage 4). This provides human auditors with a transparent, verifiable decision trace, including structured rationales, provenance links to supporting evidence, and consensus scoring, showing how the ensemble review process reduces single-model bias exposure and supports a balanced decision.

(6) Adversarial Defense and Robustness: Integrating LLMs introduces specific attack vectors, such as "prompt injection" (attempts to manipulate the model's output via malicious inputs). To address the vulnerabilities introduced by these models, the architecture includes an Input Validation Module. This module sanitizes user inputs against known adversarial patterns, specifically targeting direct and indirect prompt injection attacks (e.g., instruction override attempts), role-playing jailbreaks (attempts to force the AI out of its analytical auditor persona), and data leakage prompts (attempts to extract the underlying system prompt or sensitive organizational data).

Furthermore, to ensure the AI remains a neutral analyzer, the system enforces "system prompt hardening" through three primary methods:

- Separation of Concerns: Using structural delimiters (e.g., XML tags) to strictly isolate system instructions from untrusted user inputs.
- Instruction Ordering: Placing critical safety guardrails at the end of the prompt to leverage the LLM's recency bias.
- Structured Output Constraints: Forcing outputs into predefined formats (e.g., strict JSON or tables) to prevent the execution of malicious payloads.

Together, these techniques prevent the AI from bypassing safety protocols or generating fabricated assessments.

4.5. Analysis of LLM Consensus and Divergence in Risk Identification

The experimental evaluation presented in this work provides new empirical evidence on how state-of-the-art Large Language Models perceive and prioritize information security risks. Instead of focusing solely on accuracy, the experiment emphasized consensus and divergence in LLM-generated outputs. To quantify this, a thematic analysis methodology was applied. Human cybersecurity experts independently reviewed the raw, unstructured outputs from the five models and mapped them to standardized ISO/IEC 27005 risk domains. The primary evaluation metric for model performance in this context was the Simple Agreement Ratio (N/5), defined as the number of independent models that successfully identified a risk belonging to a specific thematic category. The results were then consolidated to reveal areas of high consensus and divergence (Table 4). While formal inter-rater reliability statistics (e.g., Fleiss' Kappa) and evaluations against expert-validated

benchmarking datasets are important for full empirical validation, they were not computed in the present architectural study. Instead, expert-led thematic coding and disagreement reconciliation were used as exploratory qualitative validation procedures. A formal k-based reliability assessment is reserved for the evaluation of the integrated prototype in future work. For the current study, the thematic agreement ratio serves as an exploratory indicator of cross-model consistency in risk identification and supports the conceptual plausibility of the proposed Multi-LLM Delphi module. However, full operational validation requires benchmark-based evaluation, formal inter-rater reliability statistics, and testing within an integrated prototype.

Table 4. Consolidated risk themes and expert consensus. The checkmark (✓) indicates that the AI model identified the risk theme, while the en dash (–) indicates it did not.

Risk Theme	ChatGPT	Gemini	Grok	Copilot	DeepSeek	Consensus
Access Control and Privilege Mgmt	✓	✓	✓	✓	✓	5/5
Insider Risk/Personnel Security	✓	✓	✓	✓	✓	5/5
Third-Party/Vendor Risk	✓	✓	✓	✓	✓	5/5
Vulnerability and Patch Mgmt	✓	✓	✓	✓	✓	5/5
Incident Response and Monitoring	✓	✓	✓	✓	✓	5/5
Data Classification and Handling	✓	✓	✓	✓	✓	5/5
Physical and Env. Security	✓	✓	✓	✓	✓	5/5
Insecure Remote Access	✓	–	✓	✓	✓	4/5
Insecure Development	–	✓	–	✓	✓	3/5
Business Continuity	–	✓	–	✓	–	2/5
Phishing/Social Engineering	✓	–	–	–	–	1/5
Denial of Service	✓	–	–	–	–	1/5
“Shadow IT” (Unauthorized Assets)	–	✓	–	–	–	1/5
Regulatory Non-Compliance	–	–	✓	–	–	1/5
System Outage (Poor Maintenance)	–	–	✓	–	–	1/5
Weak Network Segmentation	–	–	–	✓	–	1/5
Lack of Security Awareness	–	–	–	–	✓	1/5

The results reveal a high degree of convergence among all five models. Seven key themes were identified by all models (5/5 consensus), including Access Control and Privilege Management, Insider Threats, Third-Party Risk, Vulnerability and Patch Management, Incident Response and Monitoring, Data Classification and Handling, and Physical Security. These represent the core domain of shared situational understanding across LLMs and align directly with the ISO/IEC 27005 Context Establishment and Risk Identification stages.

Divergence was observed in lower-frequency or more context-dependent areas such as Insecure Development (3/5), Business Continuity (2/5), and rare single-model detections like Phishing, Shadow IT, and Weak Network Segmentation. These less frequent topics highlight model-specific analytical biases and varying domain depths, demonstrating that certain LLMs emphasize strategic governance aspects, while others focus more on operational or technical threats.

From a systems design perspective, this variation provides valuable insight for the LLM-based Decision-Making Module. By aggregating outputs from multiple models, similar to a Delphi-style consensus, the architecture can achieve more balanced, explainable, and

robust risk evaluation. Commonly detected risks can serve as baseline knowledge, while less frequent or unique detections may act as early indicators of emerging or overlooked vulnerabilities, enriching the AI's adaptive reasoning processed architecture.

5. Discussion

This section explains how the proposed architecture extends prior AI-driven risk management studies and outlines directions for further development.

5.1. Comparison with Previous Studies

This study builds upon the growing body of research exploring the application of artificial intelligence to risk management and cybersecurity automation. Similarly to previous works [13,15,16], the results of this research reaffirm that AI plays a crucial role in enhancing the efficiency and accuracy of risk management processes. Earlier studies mainly focused on demonstrating AI's potential for risk prediction, analysis, and management. However, most of them remained conceptual and did not provide a fully implementable framework.

Compared to I. Hamid & M. Rahman (2025) [13] and S. S. Dasawat & S. Sharma (2023) [16], who emphasized the strategic importance of AI adoption in cybersecurity risk management, the present work advances these concepts by proposing a structured, ISO/IEC 27005-aligned architecture that defines concrete steps for automation. N. Mohamed (2023) [15] identified that 20% of organizations hesitate to adopt AI due to a "transparency deficit." This work directly addresses this barrier via the Documented Information module (Stage 4), which ensures that every algorithmic decision generates a regulatory-compliant audit trail. Furthermore, while M. Yazdi et al. (2024) [14] explored the potential of ChatGPT-4 for risk assessment, their study revealed limitations in its precision and contextual understanding. The proposed Decision-Making System addresses these shortcomings by integrating multiple Large Language Models and applying a Delphi-style aggregation process to compare, cross-evaluate, and reconcile model-generated risk reasoning. This design may reduce reliance on any single model and broaden analytical coverage, but it does not eliminate the possibility of shared biases or guarantee correctness.

Additionally, M. Sterbak et al. (2021) [17] noted that the full automation of risk management processes remains difficult due to human-dependent subprocesses. The proposed architecture addresses this limitation more directly by combining NLP, Named Entity Recognition, and LLM-based semantic parsing to process unstructured organizational policies, system descriptions, and architectural documentation, thereby enabling automated IT asset detection, extraction, and categorization together with related threat context identification. Finally, while the current architecture effectively utilizes an ensemble of state-of-the-art general-purpose LLMs to provide a broad analytical baseline, transitioning to a domain-specific, security-oriented LLM remains a key objective for future optimization to further enhance technical accuracy and interpretability.

5.2. Limitations

It is important to acknowledge a limitation regarding the experimental evaluation of the LLMs presented in this study. The experiment employed a "Zero-Shot Role Prompting" strategy to assess the baseline, intrinsic reasoning capabilities of the models when facing a scenario without pre-labeled training data. While this approach successfully demonstrates the models' foundational understanding of ISO/IEC 27005 principles, it may oversimplify optimal real-world applications. In a practical enterprise deployment, organizations would likely utilize existing historical risk data and internal documentation to apply few-shot learning, Retrieval-Augmented Generation, or model fine-tuning. Integrating organization-

specific context through these advanced methods would undoubtedly yield more tailored, accurate, and actionable risk assessments, representing a critical optimization step for future real-world implementations.

Specifically, the consensus and divergence patterns observed in the zero-shot experiment directly inform the design of the Retrieval-Augmented Generation components within the proposed architecture. The high consensus on core risks (e.g., Access Control and Insider Threats) indicates that LLMs possess a strong foundational understanding of standard security domains. However, the observed divergence in highly contextual or lower-frequency risks (e.g., Business Continuity and Shadow IT) highlights the models' dependency on external grounding. Consequently, the architecture's context establishment (Stage 1) and case-based reasoning (Stage 3) modules function effectively as RAG pipelines. By retrieving precise organizational policies and historical incident data, these components are designed to resolve the exact types of inter-model divergence identified in our experiment. This ensures that the Multi-LLM consensus mechanism is anchored in verified corporate reality rather than relying solely on generic pre-training data, directly bridging the gap between baseline capabilities and enterprise-grade reliability.

Furthermore, as noted in Section 4.5, the exploratory experiment relied on expert-led thematic coding and a Simple Agreement Ratio rather than formal inter-rater reliability statistics, such as Fleiss' Kappa. The primary rationale for this omission relates to the specific scope and resource constraints of the current study. This paper is fundamentally an architectural design study aimed at establishing a standard-oriented framework. Conducting a rigorous, reproducible statistical assessment of LLM reliability would require a significantly larger, standardized dataset of organizational policies and a larger, blinded panel of independent expert coders to achieve statistical power. Such an endeavor aligns more closely with dedicated LLM benchmarking rather than systems architecture design. The current experiment was deliberately scoped as a small-scale proof-of-concept to assess foundational understanding and conceptually justify the inclusion of the Multi-LLM Delphi consensus mechanism. While the current findings successfully demonstrate this conceptual plausibility, the full empirical validation of the AI models' feasibility requires robust quantitative metrics. Implementing Fleiss' Kappa on a larger subset of expert-validated data will be a central methodological requirement during the future empirical evaluation of the fully integrated prototype.

A further limitation of the current conceptual architecture pertains to the computational complexity and large-scale deployment feasibility. The proposed Multi-LLM Delphi consensus mechanism inherently multiplies the computational cost and latency of inference. Specifically, for a typical risk scenario evaluated by an ensemble of five models across three phases (Independent Generation, Cross-Evaluation, and Consensus Aggregation), the system must generate over ten separate LLM calls. Consequently, the token processing cost is roughly ten times higher compared to a standard single-model query. Furthermore, even if the requests are executed simultaneously in parallel, the overall system latency is bottlenecked by the slowest model in each phase, effectively tripling the baseline inference time. In a large enterprise environment processing thousands of risk events, this would introduce significant financial overhead and processing bottlenecks.To ensure scalability, future prototype implementations will need to mitigate this overhead through a two-tiered architectural approach. First, utilizing localized, quantized open-weight models (e.g., Llama 3 and Mistral) deployed on premises will eliminate API costs and reduce network-level latency. Second, the system must integrate a robust semantic caching layer. This layer would utilize a vector database to store previous organizational policies and risk outcomes. When a new assessment request is triggered, a similarity search will compare the incoming context against the cache. If a high-confidence match is found for a routine or recurring

threat, the system will retrieve the cached consensus, entirely bypassing the expensive Multi-LLM engine. The development of this functional software prototype, along with empirical computational benchmarking of the caching layer in a live environment, remains a primary objective for the subsequent research phase.

A further limitation concerns the interpretation of multi-LLM consensus. Although the Delphi-style aggregation mechanism can diversify perspectives and reduce reliance on any single model output, it should not be interpreted as a guarantee of objectivity, bias elimination, or correctness. Large language models may share systemic biases, overlapping training data influences, or similar alignment tendencies, meaning that consensus can still reflect correlated error rather than true validity. Accordingly, in the proposed architecture, multi-model agreement is treated as a decision support signal that must remain subject to human review and organizational context

5.3. Future Work

This study represents the first stage in a series of works focused on the automation of information security risk management. It establishes the foundational principles for future research and introduces the proposed architecture, which outlines the core process stages.

While prior versions of ISRM models listed reinforcement learning generically, in our proposed architecture, RL is treated as an offline, optional component with a clearly defined organizational reward and explicit data sources for ground truth (as detailed in Section 4.2). The primary decision engine remains strictly auditable (constrained optimization combined with case-based reasoning). RL is proposed specifically to learn treatment sequencing over time. However, as no empirical validation of the RL agent was conducted in the current study, this component remains theoretical. The empirical validation of this offline RL module, including its actual training on historical organizational logs, simulator calibration, and off-policy evaluation, is explicitly positioned as a primary objective for future work. This planned empirical phase will address data scarcity and safety concerns while preserving the learning benefits for recurrent cybersecurity workflows.

The next research article will focus on the development and evaluation of NLP methods designed to process internal organizational documents and to integrate data from the MEHARI project outputs. A subsequent study will address the implementation of the Decision-Making System, including the testing of a custom Large Language Model specialized in information security contexts.

As organizations increasingly adopt AI, the risk landscape itself evolves. Future iterations of the proposed architecture will extend the cross-model agreement analysis to include AI-specific threat domains, such as those defined in the MITRE ATLAS (Adversarial Threat Landscape for Artificial Intelligence Systems) framework. Incorporating ATLAS will enable the system to identify risks unique to AI deployments, such as data poisoning and model evasion, ensuring the risk management process remains resilient against next-generation threats.

Finally, the concluding paper in this research cycle will present the developed prototype of the semi-automated risk management system along with the experimental validation of results and performance evaluation metrics. These planned studies are expected to contribute toward the realization of a semi-automated, AI-powered information security risk management system that aligns with international standards and practical cybersecurity needs.

6. Conclusions

The conducted research confirmed the growing importance of artificial intelligence in the field of ISRM. The analysis of the degree of research development has shown

that studies in this domain are highly relevant and promising, though their practical implementation remains challenging. Previous works have demonstrated AI's potential in enhancing decision making, risk prediction, and automation but have lacked unified, standard-oriented architectures.

The comparative analysis of eight international standards revealed notable differences in their suitability for automation. While NIST SP 800-53 (4.20) and ISO 31000 (4.00) demonstrated strong potential in compliance verification and high-level governance respectively, they lack the specific procedural flow required for autonomous decision making. Frameworks such as COSO IC-IF (3.20) and IEC 62443-2-1 (3.40) remained focused on organizational or industrial niches, limiting their broader algorithmic applicability. In contrast, ISO/IEC 27005 achieved the highest final score (4.60), distinguishing itself through superior Automation Readiness and process decomposition. Its balance between theoretical rigor and operational guidance makes it the optimal foundation for AI-driven automation, enabling the direct mapping of risk management steps (context establishment, assessment, treatment, and documentation) to ML and NLP modules.

The scientific novelty of this work lies in the proposed architecture integrating NLP, Big Data Analytics, and machine learning techniques for automated risk identification, assessment, and treatment. The innovative Decision-Making System, combining expert reasoning through the Delphi method with the analytical capabilities of LLMs, establishes a new direction for intelligent and adaptive IS management.

The practical significance of the study is reflected in the design of an architecture that can be integrated into governmental and corporate information systems, enabling more consistent, data-driven, and efficient risk management. This research establishes ISO/IEC 27005 as the cornerstone for AI-powered risk management automation. Rather than claiming a fully autonomous solution at this stage, the study presents a theoretical reference architecture that maps standard-defined processes to specific AI components. This groundwork enables the future development of semi-automated ISRM systems where human oversight is augmented by algorithmic consistency, aligning with international standards and practical cybersecurity needs.

As an exploratory assessment supporting the proposed approach, an experimental comparison of five state-of-the-art Large Language Models was conducted using a standardized ISO/IEC 27005 risk identification task. Instead of measuring accuracy against a fixed benchmark, the experiment analyzed the level of consensus and divergence among models in identifying and prioritizing organizational risks. The results revealed strong cross-model agreement in core domains such as access control, insider threats, and incident response, while differences emerged in less-defined areas like business continuity and social engineering. These findings provide exploratory support for the view that LLMs may contribute to expert-informed risk identification and structured decision support within the proposed ISO/IEC 27005-based architecture, particularly within the LLM-driven Decision-Making System designed for consensus-based risk evaluation; however, they do not yet establish the full operational feasibility of the architecture, which requires future prototype-based validation.

The main contributions of this study correspond to the three research questions:

- C1: The comparative evaluation identified ISO/IEC 27005 as the most suitable standard for AI-based automation due to its structured, detailed, and process-oriented design.
- C2: Each ISO/IEC 27005 stage was linked to specific AI methods: NLP for document analysis, ML for risk prediction, and Big Data for threat intelligence.
- C3: The proposed AI-driven architecture is designed to enhance risk management by supporting analytical consistency, structured decision support, and process automation through expert-informed LLM reasoning.

In summary, the presented study provides a theoretical foundation and exploratory support for AI-assisted automation in risk management, while the practical feasibility of the full architecture remains to be established through future prototype-based validation.

Author Contributions: Conceptualization, O.C. and K.D.; methodology, O.C.; software, O.C.; validation, O.C., K.D., Š.G. and R.B.; formal analysis, O.C.; investigation, O.C.; resources, O.C. and K.D.; data curation, O.C.; writing—original draft preparation, O.C.; writing—review and editing, O.C., K.D. and Š.G.; visualization, O.C.; supervision, K.D.; project administration, O.C.; and funding acquisition, K.D., Š.G. and R.B. All authors have read and agreed to the published version of the manuscript.

Funding: This paper was funded by the project "Research on Cyber Resilience Through Application of Generative Artificial Intelligence in Chief Information Security Officer Operations", which has received funding from the Research Council of Lithuania (LMTLT) (agreement No S-ITP-24-13).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this paper are available on request from the corresponding author. The incompleteness of the project prevents the data from being publicly available.

Acknowledgments: The authors thank *Electronics* for the opportunity to publish and share this research in their journal. During the preparation of this manuscript, the authors used the ChatGPT (GPT-4o model) web application for the purposes of language editing and grammar refinement. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Siponen, M.T.; Oinas-Kukkonen, H. A review of information security issues and respective research contributions. *ACM SIGMIS Database* **2007**, *38*, 60–80. [CrossRef]
2. Lundgren, B.; Möller, N. Defining Information Security. *Sci. Eng. Ethics* **2019**, *25*, 419–441. [CrossRef] [PubMed]
3. Borky, J.M.; Bradley, T.H. Protecting Information with Cybersecurity. In *Effective Model-Based Systems Engineering*; Springer: Cham, Switzerland, 2019; pp. 123–135. [CrossRef]
4. Clinton, L.; Todt, K. *Fixing American Cybersecurity: Creating a Strategic Public-Private Partnership*; Georgetown University Press: Washington, DC, USA, 2023.
5. Bhattacharjee, J.; Sengupta, A.; Mazumdar, C.; Barik, M.S. A two-phase quantitative methodology for enterprise information security risk analysis. In Proceedings of the CUBE International Information Technology Conference, Pune, India, 3–5 September 2012; pp. 809–815. [CrossRef]
6. Allodi, L.; Massacci, F. Security events and vulnerability data for cybersecurity risk estimation. *Risk Anal.* **2017**, *37*, 1606–1627. [CrossRef] [PubMed]
7. Jarrahi, M.H. Artificial Intelligence and the Future of work: Human-AI Symbiosis in Organizational Decision Making. *Bus. Horiz.* **2018**, *61*, 577–586. [CrossRef]
8. Chalyi, O. An evaluation of general-purpose AI chatbots: A comprehensive comparative analysis. *InfoSci. Trends* **2024**, *1*, 52–66. [CrossRef]
9. Yenduri, G.; Murugan, R.; Maddikunta, P.K.R.; Bhattacharya, S.; Sudheer, D.; Savarala, B.B. Artificial General Intelligence: Advancements, Challenges, and Future Directions in AGI Research. *IEEE Access* **2025**, *13*, 134325–134356. [CrossRef]
10. Jabbar, H.; Al-Janabi, S.; Syms, F. AI-Integrated Cyber Security Risk Management Framework for IT Projects. In *2024 International Jordanian Cybersecurity Conference (IJCC)*; IEEE: Piscataway, NJ, USA, 2024; pp. 76–81. [CrossRef]
11. Stahl, B.C. *Artificial Intelligence for a Better Future*; Springer International Publishing: Cham, Switzerland, 2021. [CrossRef]
12. *ISO/IEC 27005:2022*; Information Security, Cybersecurity and Privacy Protection—Guidance on Managing Information Security Risks. ISO: Geneva, Switzerland, 2022.
13. Hamid, I.; Rahman, M.M.H. AI, machine learning and deep learning in cyber risk management. *Discov. Sustain.* **2025**, *6*, 389 . [CrossRef]
14. Yazdi, M.; Zarei, E.; Adumene, S.; Beheshti, A. Navigating the Power of Artificial Intelligence in Risk Management: A Comparative Analysis. *Safety* **2024**, *10*, 42. [CrossRef]

15. Mohamed, N. Current Trends in AI and ML for cybersecurity: A state-of-the-art Survey. *Cogent Eng.* **2023**, *10*, 2272358. [CrossRef]
16. Dasawat, S.S.; Sharma, S. Cyber Security Integration with Smart New Age Sustainable Startup Business, Risk Management, Automation and Scaling System for Entrepreneurs: An Artificial Intelligence Approach. In Proceedings of the 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 17–19 May 2023; pp. 1357–1363. [CrossRef]
17. Sterbak, M.; Segec, P.; Jurc, J. Automation of risk management processes. In Proceedings of the 2021 19th International Conference on Emerging eLearning Technologies and Applications (ICETA), Starý Smokovec, Slovakia, 11–12 November 2021. [CrossRef]
18. Shamala, P.; Ahmad, R.; Zolait, A.; Sedek, M. Integrating information quality dimensions into information security risk management (ISRM). *J. Inf. Secur. Appl.* **2017**, *36*, 1–10. [CrossRef]
19. *ISO/IEC 13335-1:2004*; Information Technology—Security Techniques—Management of Information and Communications Technology Security—Part 1: Concepts and Models for Information and Communications Technology Security Management. ISO: Geneva, Switzerland, 2004.
20. ISACA. *COBIT 2019 Framework: Governance and Management Objectives*; ISACA: Schaumburg, IL, USA, 2018.
21. *ISO 31000:2018*; Risk Management—Guidelines. ISO: Geneva, Switzerland, 2018.
22. Information Security Forum. *The Standard of Good Practice for Information Security*; Information Security Forum: London, UK, 2024.
23. Joint Task Force Interagency Working Group. *Security and Privacy Controls for Information Systems and Organizations*; NIST Special Publication 800-53, Revision 5; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2020.
24. *IEC 62443-2-1:2010*; Industrial Communication Networks—Network and System Security—Part 2-1: Establishing an Industrial Automation and Control System Security Program. IEC: Geneva, Switzerland, 2010.
25. Committee of Sponsoring Organizations of the Treadway Commission (COSO). *Internal Control—Integrated Framework: Executive Summary*; COSO: Durham, NC, USA, 2013.
26. HITRUST Alliance. *HITRUST Risk Management Handbook*; HITRUST Alliance: Frisco, TX, USA, 2023.
27. Center for Internet Security. Information Security Policy Template. Available online: https://www.cisecurity.org/-/media/project/cisecurity/cisecurity/data/media/files/uploads/2020/06/Information-Security-Policy.docx (accessed on 30 October 2025).
28. *ISO/IEC 27001:2022*; Information Security, Cybersecurity and Privacy Protection—Information Security Management Systems—Requirements. ISO: Geneva, Switzerland, 2022.
29. Chalyi, O. Assessing Wi-Fi Security Protocols: A Study of Dictionary Attack Performance. *Balt. J. Mod. Comput.* **2025**, *13*, 592–623. [CrossRef]
30. Rivai, M.A.; Suroso, J.S.; Pangemanan, F. Review of the Risk Analysis Using MEHARI Model: The Guideline to Analyze Risk for Startup Educational Platform. In Proceedings of the 2020 International Conference on Information Management and Technology (ICIMTech), Bandung, Indonesia, 13–14 August 2020; pp. 1–6. [CrossRef]
31. *ISO/IEC 27002:2022*; Information Security, Cybersecurity and Privacy Protection—Information Security Controls. ISO: Geneva, Switzerland, 2022.
32. Chalyi, O.; Driaunys, K.; Rudžionis, V. Assessing Browser Security: A Detailed Study Based on CVE Metrics. *Future Internet* **2025**, *17*, 104. [CrossRef]