HAKEMLİ Denetişim
Ortak Aklın Harmanı
Kamu İç Denetçileri Derneği Meşrutiyet Caddesi Konur Sokak No: 36/6 Kızılay - ANKARA
www.kidder.org.tr/denetisim/ • denetisim@kidder.org.tr
ISSN 1308-8335
Yıl: 17, Sayı: 35, 418-437, 2026
Gönderilme Tarihi: 9 Haziran 2026 Kabul Tarihi: 6 Ağustos 2026

Research Article

AUDITING ARTIFICIAL INTELLIGENCE

(YAPAY ZEKÂ DENETİMİ)

Osman TURAN[1], Zümrüt MÜFTÜOĞLU[2], Tolga ÖZBİLGE[3]

ABSTRACT

The rapid proliferation of artificial intelligence (AI) systems across critical sectors has introduced significant security, privacy, and ethical challenges. Traditional information security audit frameworks remain insufficient to address the unique risks associated with AI technologies, particularly in areas such as data integrity, model robustness, and algorithmic transparency. This study proposes a comprehensive and risk-based AI audit methodology that integrates governance, data security, model security, and application security dimensions into a unified control framework. The proposed methodology is structured around key control domains, including documentation, compliance, access control, application security, and model security, supported by representative control considerations. The proposed methodology supports systematic evaluation of AI audit controls through a structured assurance-oriented framework. Additionally, a risk prioritization approach is introduced to classify controls into different criticality levels, enabling organizations to focus on high-impact vulnerabilities such as data poisoning, model inversion, prompt injection, and sensitive data leakage. The methodology is designed as a step-by-step audit process, including scope definition, control mapping, evidence collection, evaluation, and reporting. This structured approach ensures both technical and organizational aspects of AI systems are systematically assessed. The study contributes to the literature by providing a practical, measurable, and adaptable framework that aligns with regulatory requirements such as data protection laws and ethical AI principles. Overall, the proposed AI audit framework enhances the auditability, transparency, and security of AI systems, offering organizations a robust tool to manage emerging AI-related risks effectively.

Keywords: AI Audit, AI Security, Risk-Based Audit, AI Governance, Compliance

JEL Classification: *M42*

ÖZ

Yapay zeka (YZ) sistemlerinin kritik sektörlerde hızla yaygınlaşması, güvenlik, gizlilik ve etik açıdan önemli riskleri beraberinde getirmektedir. Geleneksel bilgi güvenliği denetim yaklaşımları, veri bütünlüğü, model güvenliği ve algoritmik şeffaflık gibi yapay zekaya özgü riskleri ele almakta yetersiz kalmaktadır. Bu çalışma, yönetişim, veri güvenliği, model güvenliği ve uygulama güvenliği boyutlarını kapsayan bütüncül ve risk tabanlı bir yapay zeka denetim metodolojisi önermektedir. Önerilen metodoloji, temsili kontrol hususlarıyla desteklenen dokümantasyon, uyumluluk, erişim kontrolü, uygulama güvenliği ve model güvenliği dahil olmak üzere temel kontrol alanları etrafında yapılandırılmıştır. Önerilen metodoloji, yapılandırılmış, güvence odaklı bir çerçeve aracılığıyla yapay zeka denetim kontrollerinin sistematik değerlendirilmesini desteklemektedir. Ayrıca, kontroller risk seviyelerine göre önceliklendirilerek veri zehirleme, model tersine mühendislik, prompt enjeksiyonu ve hassas veri sızıntısı gibi kritik tehditlere odaklanılması sağlanmaktadır. Metodoloji; kapsam belirleme, kontrol eşleştirme, veri toplama, değerlendirme ve raporlama adımlarından oluşan sistematik bir denetim süreci sunmaktadır. Bu yapı sayesinde yapay zeka sistemlerinin hem teknik hem de yönetsel boyutları bütüncül bir şekilde değerlendirilebilmektedir. Çalışma, ölçülebilir, uygulanabilir ve düzenleyici gereksinimlerle uyumlu bir çerçeve sunarak literatüre katkı sağlamaktadır. Sonuç olarak önerilen yapay zeka denetim çerçevesi, YZ sistemlerinin denetlenebilirliğini, şeffaflığını ve güvenliğini artırarak kurumların ortaya çıkan riskleri etkin bir şekilde yönetmesine yardımcı olmaktadır.

Anahtar Kelimeler: *Yapay Zeka Denetimi, Yapay Zeka Güvenliği, Risk Tabanlı Denetim, Yapay Zeka Yönetişimi, Uyum*

JEL Kodları: *M42*

[1] Expert, Social Sciences University of Ankara, Orcid Id: 0009-0005-9146-4533, osman.turan@student.asbu.edu.tr
[2] Expert, Turkish Standardization Institute, Ankara, Orcid Id: 0000-0003-3754-749, zumrutmuftuoglu@gmail.com
[3] Auditor, Min. of Lab. and Soc. Security, Ankara, Orcid Id:0009-0004-0368-3545, tolgaozbilge@yahoo.com

1. INTRODUCTION

Artificial Intelligence (AI) technologies have been extensively adopted across numerous critical sectors, including finance, healthcare, defense, public administration, and manufacturing (Acemoglu & Restrepo, 2019). In particular, advancements in Machine Learning (ML) and Deep Learning-based systems have enabled the processing of vast datasets and the automation of complex decision-making processes (Jordan & Mitchell, 2015; LeCun et al., 2015). However, this rapid proliferation has introduced novel risks concerning security, privacy, ethics, and governance (Amodei et al., 2016; Jobin et al., 2019).

Traditional information security paradigms are primarily constructed upon static systems and deterministic software components (Anderson, 2020). Conversely, AI systems—characterized by data dependency, learning capabilities, unpredictable model behaviors, and dynamically updatable structures—cannot be fully evaluated through classical auditing methods (Falco et al., 2021; Raji et al., 2020). Specifically, the security of training datasets, model integrity, algorithmic bias, and explainability have necessitated audit requirements unique to AI systems (Papernot et al., 2018; Floridi et al., 2018).

Various frameworks have been proposed in the literature for AI security and governance. Frameworks such as the NIST AI Risk Management Framework (NIST, 2023), the ISO/IEC 23894 standard, and the OWASP Top 10 for LLM Applications (OWASP, 2023) provide significant contributions toward defining and managing AI-related risks. Nevertheless, most of these approaches offer high-level principles and remain limited in providing detailed, measurable audit controls at the implementation level (Schiff et al., 2020).

Although recent frameworks such as the NIST AI Risk Management Framework, ISO/IEC 42001, ISO/IEC 23894, and the OWASP Top 10 for LLM Applications have substantially advanced AI governance and security practices, they primarily focus on high-level principles, risk management processes, or specific technical vulnerabilities. Consequently, organizations often face difficulties in translating these recommendations into a coherent and auditable control structure that can be consistently applied throughout the AI lifecycle. Existing approaches frequently address governance, data protection, application security, and model security independently, resulting in fragmented audit activities and inconsistent assurance practices. This fragmentation motivates the need for an integrated audit framework capable of combining organizational governance with AI-specific technical controls within a single, risk-oriented methodology.

The proposed framework is designed around four conceptual dimensions that collectively represent the essential pillars of AI auditability: AI governance, data security, application security, and model security. These conceptual dimensions are operationalized through five complementary control domains. Governance is implemented through documentation and compliance controls, while data security is primarily addressed through access and authorization controls. Application security and model security remain independent technical domains because they require distinct security objectives, evaluation procedures, and assurance evidence. This distinction enables the framework to preserve conceptual simplicity while providing operational granularity required for practical audit execution.

In this context, concrete, actionable, and measurable control mechanisms are required for organizations to effectively audit AI systems (Brundage et al., 2020). A holistic auditing approach—integrating data security, model security, and application security—is of critical importance in managing AI-induced risks.

Aiming to address this gap, this study proposes a comprehensive, risk-based audit methodology for AI systems. The proposed methodology is based on control objectives encompassing core domains such as documentation, compliance, access control, application security, and model security. Furthermore, it is supported by a scoring model that facilitates the evaluation of each control based on its maturity level (Metcalf et al., 2021).

The primary contributions of this study are summarized as follows:

- Systematic Classification: A systematic classification of security and audit requirements specific to AI systems.
- Holistic Framework: Provision of an integrated audit framework that converges technical and administrative controls (Kroll et al., 2015).
- Maturity-Based Evaluation: Development of a maturity-based assessment model that ensures the measurability of controls (Caralli et al., 2010).

- Risk-Based Prioritization: Focus on critical vulnerabilities through a prioritization approach based on risk levels.

In conclusion, this study aims to contribute to the secure, transparent, and auditable development and operation of AI systems. The proposed methodology is expected to serve as a significant reference framework for both academic literature and practical applications (Cath, 2018; Gasser & Almeida, 2017).

2. LITERATURE REVIEW

The rapid integration of artificial intelligence (AI) systems into critical decision-making processes has intensified global concerns regarding transparency, fairness, and accountability (Jobin et al., 2019; Mökander, 2023). Unlike traditional information systems, machine learning models exhibit complex, data-driven, and often opaque behaviors that challenge conventional auditing practices and necessitate a fundamental shift in governance (Burrell, 2016; Leocádio et al., 2024). Within this context, Mökander and Floridi (2021) argue that ethics-based auditing serves as a vital bridge to translate abstract principles into practical AI governance, ultimately improving decision-making and reducing human suffering. They suggest that for auditing to be effective, it must be a continuous, constructive, and systemic process rather than a one-time technical check, a sentiment echoed by Raji et al. (2020), who conceptualize AI auditing as a multidimensional evaluation encompassing system performance, organizational processes, and societal impact. By aligning these audits with public policy and market incentives, organizations can potentially unlock economic growth while ensuring their systems remain trustworthy; however, this transition is currently hindered by significant conceptual, technical, and economic constraints (Mökander & Floridi, 2021).

The existing literature on AI auditing reveals three dominant yet largely fragmented perspectives that frequently operate in isolation, leading to incomplete oversight. Technical auditing approaches focus on measurable properties such as model accuracy, robustness, and bias detection (Mitchell et al., 2019; Goodfellow et al., 2015), yet these quantitative insights are often limited to model-level evaluation and fail to capture broader socio-technical implications (Selbst et al., 2019; Biggio & Roli, 2018). Process-oriented auditing extends the scope of analysis to the entire AI lifecycle, including data collection and deployment (Costanza-Chock, 2020; Leocádio et al., 2024), while ethical and legal approaches assess systems in relation to normative principles such as non-maleficence and regulatory requirements (Floridi et al., 2018). To bridge the gap between these aspirational principles and enforceable standards, Falco et al. (2021) propose a framework built upon the governance principles: prospective risk assessments to identify harms before deployment, operational audit trails to ensure traceability during use, and system adherence to jurisdictional requirements. By shifting the focus to these tangible pillars, independent auditing transforms burdensome ethical guidance into an enforceable and scalable safety standard, addressing the accountability gap highlighted by high-profile AI failures in autonomous driving and medical diagnostics.

A central concept that has gained prominence in recent research is AI auditability, defined as the extent to which an AI system can be effectively audited (Brundage et al., 2020). Auditability is not an inherent characteristic of AI systems but rather a property that must be intentionally designed and implemented through transparency, explainability, traceability, and reproducibility (Guidotti et al., 2018). Recent studies emphasize a paradigm shift toward “auditability-by-design,” where auditing capabilities are embedded throughout the AI lifecycle rather than applied retrospectively. This shift aligns closely with software engineering practices, where principles such as version control and continuous integration are adapted to support "AuditOps," reflecting the integration of auditing into the operational lifecycle of AI systems. This becomes particularly critical when addressing "frontier AI"—highly capable foundation models that may develop dangerous, unexpected capabilities (Anderljung et al., 2023). Anderljung et al. (2023) argue that traditional regulatory methods are insufficient because model risks can emerge suddenly, advocating for a safety-first approach that includes mandatory pre-deployment risk assessments, external auditing, and continuous post-deployment monitoring.

The governance of these frontier models requires a departure from standard IT auditing due to the stochastic nature of large-scale neural networks. Unlike deterministic software, AI behavior is often emergent, meaning that rigorous safety standards must be coupled with registration and reporting requirements for transparency (Anderljung et al., 2023). This suggests that while industry self-regulation is a starting point, government intervention, such as licensing regimes and supervisory authorities, is necessary to protect public safety. The literature underscores that the complexity of these models necessitates independent audits as a pragmatic, multidisciplinary mechanism for widespread assurance (Falco et al., 2021). Without such independent verification, the "black box" nature of proprietary models and restricted dataset access will continue to hinder external auditing efforts and degrade public trust (Sandvig et al., 2014; Birhane et al., 2024).

Furthermore, the expansion of AI auditing tools has not yet resolved the fundamental issues of interoperability and end-to-end lifecycle support. Existing tools are predominantly evaluation-centric, focusing on fairness metrics and bias detection rather than providing a cohesive accountability infrastructure that integrates technical tools with organizational processes (Metaxa et al., 2021). This fragmentation is compounded by the human and organizational dimensions of AI auditing. Traditional IT auditors often lack the necessary machine learning and ethical expertise required to evaluate modern AI. Consequently, AI auditing is increasingly viewed as a multidisciplinary endeavor involving collaboration among data scientists, engineers, auditors, legal experts, and social scientists (Raji et al., 2022). This shift is further reflected in the introduction of frameworks based on the capability approach, which evaluate AI systems in terms of their impact on human freedoms and well-being (Nussbaum, 2009).

Recent perspectives also emphasize the importance of micro ethical auditing, suggesting that ethical responsibility should be embedded within everyday development practices—such as data selection and feature engineering—rather than being confined to external, retrospective reviews. This redefines the audit not just as a compliance check but as a continuous and participatory process where developers play an active role in risk mitigation. Moreover, the literature highlights the necessity of aligning AI auditing with stakeholder needs. AI auditing is not only a control mechanism but also a knowledge-generating process that provides guidance and awareness to developers, organizations, and end-users (Ananny & Crawford, 2018). However, the lack of standardized definitions of ethical principles and inconsistencies in their operationalization remain significant barriers (Wieringa, 2020).

The organizational infrastructure supporting AI auditing must also evolve to accommodate the "AuditOps" paradigm. This involves the creation of standardized documentation practices, such as model cards and datasheet for datasets, which facilitate traceability and reproducibility throughout the system's existence (Mitchell et al., 2019; Gebru et al., 2021). Such documentation serves as the "audit trail" described by Falco et al. (2021), ensuring that auditors can reconstruct the decision-making history of a model when a failure occurs. Furthermore, the integration of continuous monitoring systems allows for the detection of "model drift" or performance degradation in real-world environments, which is essential for ensuring long-term non-maleficence (Floridi et al., 2018).

Despite these advancements, the field is characterized by a significant gap between theory and practice. While many organizations subscribe to high-level ethical principles, the practical implementation of these principles through auditing is often hampered by the costs associated with deep technical evaluations and the competitive pressure to deploy models quickly (Mökander & Floridi, 2021). The weak linkage between audit findings and actionable outcomes often means that even when risks are identified, there are no clear mechanisms for remediation or enforcement (Wieringa, 2020). Future research must therefore focus on the development of holistic, interoperable, and actionable AI auditing frameworks that can bridge the gap between technical metrics and societal values. This requires not only technical innovation in bias detection and explainability but also the development of institutional frameworks that empower auditors to hold AI developers accountable for the societal impact of their systems (Cobbe et al., 2020).

The literature demonstrates that AI auditing has evolved considerably during the last decade. Existing studies have addressed important aspects such as AI governance, ethical auditing, transparency, auditability-by-design, continuous monitoring, adversarial robustness, and lifecycle management. Likewise, international standards and frameworks, including the NIST AI Risk Management Framework, ISO/IEC 42001, ISO/IEC 23894, and the OWASP Top 10 for LLM Applications, have established valuable principles for managing AI-related risks.

Despite these advances, several important research gaps remain. First, existing approaches largely address governance, technical security, compliance, and model evaluation as separate disciplines. Consequently, organizations often rely on multiple disconnected frameworks, resulting in fragmented audit activities and inconsistent evaluation methodologies. Second, although current standards provide comprehensive governance principles, they generally offer limited implementation-oriented guidance regarding measurable audit controls that can be systematically verified during audit engagements. Third, most existing studies focus on specific phases of the AI lifecycle or particular categories of AI risks rather than providing an integrated control structure covering governance, data management, application security, and model security simultaneously. Finally, while recent literature increasingly emphasizes trustworthy AI and continuous assurance, very few studies propose a structured assurance model that determines the depth of audit activities according to the risk profile of AI systems.

These observations directly motivated the design of the proposed framework. Rather than introducing another high-level governance model, this study translates the findings of the literature into a practical audit architecture consisting of operational control domains, measurable control objectives, and risk-oriented assurance activities. The framework integrates organizational governance requirements with AI-specific technical security controls, enabling auditors to evaluate AI systems through a single, coherent methodology.

Furthermore, the proposed Artificial Intelligence Assurance Level model extends existing AI governance approaches by introducing a graduated assurance structure inspired by assurance-oriented security evaluation methodologies. Instead of treating all AI systems equally, the proposed model scales verification activities according to system criticality and risk exposure, thereby supporting proportional and evidence-based AI auditing. Gap analyses and research motivation are illustrated in Table 1.

In conclusion, the reviewed literature suggests that AI auditing is transitioning from a fragmented set of practices into a comprehensive socio-technical discipline. Effective AI auditing requires the integration of technical auditability, organizational infrastructure, human expertise, and ethical awareness. The field remains in a formative stage, characterized by limited empirical validation and a lack of standardized frameworks (Birhane et al., 2024). Addressing these gaps is essential for the trustworthy and accountable deployment of AI systems across all sectors of society. By moving toward a systemic, continuous, and independent auditing model, the AI community can ensure that automation serves the public good while minimizing the risks of harm (Falco et al., 2021; Anderljung et al., 2023). This evolution from simple verification to holistic governance represents the most promising path forward for managing the complex risks inherent in modern artificial intelligence.

Table 1. Gap Analysis and Research Motivation

Limitation Identified in Studies	Literature	Proposed Framework Component
Fragmented AI auditing approaches	Raji et al. (2020), Falco et al. (2021)	Unified AI audit framework
High-level governance principles with limited implementation guidance	NIST AI RMF, ISO/IEC 42001, ISO/IEC 23894	Structured operational control framework
Limited integration between governance and technical security	Mökander & Floridi (2021), Leocádio et al. (2024)	Integrated governance and technical control architecture
Insufficient lifecycle-oriented audit support	Leocádio et al. (2024), Anderljung et al. (2023)	Lifecycle-based control domains
Lack of scalable assurance methodology	Existing AI governance literature	AI assurance levels
Limited support for AI-specific threats	OWASP (2023), Papernot et al. (2018)	Model security and application security controls
Weak linkage between risks and audit evidence	Wieringa (2020), Cobbe et al. (2020)	Risk-based control prioritization and assurance evidence

3. AI AUDIT: ANALYSIS, PROCEDURAL FRAMEWORK, AND DISCUSSION

The auditing of Artificial Intelligence (AI) systems possesses a multi-layered, dynamic, and highly uncertain structure that distinguishes it from traditional Information Technology (IT) auditing. This complexity necessitates the development of new methodological approaches from both technical and governance perspectives (Mökander & Floridi, 2021). Principal challenges encountered in AI auditing include the lack of standardized audit frameworks, the inherent high complexity of systems, and the fact that these systems are frequently developed by third parties (Raji et al., 2020). Consequently, AI auditing should be treated not merely as a technical examination but as a holistic evaluation process encompassing organizational, ethical, and legal dimensions.

The complexity of AI systems serves as a significant source of uncertainty during the audit process. These systems are constituted through the integration of data warehouses, machine learning algorithms, cloud computing infrastructures, and various software components (Falco et al., 2021). Therefore, rather than solely examining algorithmic accuracy, auditors should focus on governance mechanisms covering the entire system lifecycle. Although the "black box" effect attributed to machine learning models is frequently emphasized in audit processes, this does not necessarily mandate that auditors perform direct algorithmic analysis. Instead, the auditing approach should concentrate on control mechanisms, data governance, system integration, and risk management.

The starting point for an AI audit is the clear definition of the audit scope and objectives. In this stage, assessing the risks associated with the organization’s use of AI is of critical importance. To address risks systematically, tools such as the Risk and Control Matrix (RCM) should be utilized, and each risk must be documented alongside its corresponding

control mechanisms (COSO, 2023). This approach enhances the traceability of the audit process and facilitates a systematic evaluation.

Data management plays a pivotal role in the effectiveness of the audit process. The accuracy and reliability of AI systems depend largely on the quality of the datasets utilized (Sambasivan et al., 2021). Therefore, the accuracy, integrity, timeliness, and representativeness of data sources must be examined in detail within the scope of the audit. Furthermore, data retention policies must align with business requirements and legal regulations. The existence of control mechanisms applied throughout the data lifecycle directly impacts audit reliability.

Another significant dimension of AI auditing involves ethical and legal evaluations. In particular, data bias stands out as one of the most critical risk areas for AI systems. Historical or structural biases present in training data can lead to discrimination in model outputs (Mehrabi et al., 2021). Consequently, the audit process must analyze not only model performance but also the model's impact on different data groups. Bias analyses should be conducted using metrics such as fairness criteria, risk disparity, and equal opportunity (Barocas et al., 2023). Additionally, in accordance with the principles of algorithmic transparency and accountability, system explainability must be ensured.

Transparency emerges as a fundamental requirement in the auditing of AI systems. The audit process should be conducted within an iterative structure that supports continuous improvement throughout the system lifecycle (Jobin et al., 2019). In this context, model development processes, the datasets used, and the decisions made must be clearly documented. Transparency is not only a technical requirement but also a critical element for establishing stakeholder trust.

Stakeholder participation also plays a vital role in the audit process. Since AI systems are developed through the contributions of multi-disciplinary teams, technical teams, business units, legal advisors, and third-party providers should be included in the audit process (Costanza-Chock, 2020). The proliferation of cloud-based solutions, in particular, increases vendor risks and complicates audit procedures. Therefore, supply chain security and third-party risk management should be treated as integral parts of the AI audit.

Technical verification and model performance evaluation are also of critical importance. Both white-box and black-box testing approaches should be employed in model validation processes; performance metrics such as accuracy, precision, and consistency must be analyzed (Ashmore et al., 2021). Furthermore, the system's robustness and generalization capacity should be measured by evaluating how the model behaves under different data scenarios. For systems characterized by continuous learning, changes in model performance over time must be monitored, and retraining processes should be implemented when necessary.

The security dimension is another critical component of AI auditing. With the widespread adoption of Generative AI (GenAI) systems, data privacy, model security, and cyber threats have become increasingly prominent (Anderljung et al., 2023). The potential for data entered into GenAI systems to be accessed by third parties poses a risk of leaking sensitive information. Moreover, the outputs generated by these systems may not always be reliable, potentially leading to erroneous results known as "hallucinations" (Ji et al., 2023). Therefore, explicit policies regarding the use of AI systems must be established, and users should be sensitized to the risks associated with these systems.

In conclusion, AI auditing requires a multidisciplinary approach encompassing technical, organizational, and ethical dimensions. For an effective audit process, it is essential to accurately define the scope, adopt a risk-based approach, strengthen data management, and implement transparency principles.

4. THE PROPOSED APPROACH: AI AUDIT AND ASSURANCE FRAMEWORK

4.1. Research Approach

This study adopts a qualitative and design-oriented research methodology to establish a comprehensive control framework for auditing artificial intelligence systems. The research is conducted within the paradigm of Design Science Research (DSR), which aims to develop novel model proposals by systematically analyzing existing knowledge (Hevner et al., 2004).

Throughout the research process, national and international standards, academic studies, and sectoral best practices in the fields of AI security, information security management, and the software development life cycle (SDLC) were examined. The literature review focused on identifying current and high-impact sources, with particular emphasis on studies covering risk management, data security, model security, and ethical compliance (Mökander & Floridi, 2021).

The data gathered through this process were evaluated using content analysis, from which control requirements aimed at enhancing the auditability of AI systems were systematically derived. The findings were structured to provide a holistic and applicable auditing approach (Raji et al., 2020).

In addition, the research draws upon the principles of the Common Criteria for Information Technology Security Evaluation (ISO/IEC 15408), which provides a structured assurance framework for assessing the security properties of information technology products and systems. The Common Criteria approach was particularly valuable in informing the development of auditable control requirements, as it emphasizes systematic security functional requirements (SFRs), security assurance requirements (SARs), and evidence-based evaluation processes. By adapting assurance-oriented concepts from Common Criteria to the context of artificial intelligence systems, the proposed framework seeks to strengthen transparency, traceability, and verifiability throughout the AI system lifecycle, thereby enhancing the overall effectiveness and reliability of AI audits (Common Criteria Recognition Arrangement [CCRA], 2022; ISO/IEC 15408-1:2022).

The proposed framework was developed through an iterative thematic analysis of the reviewed literature, international standards, and industry guidance documents. The literature review and qualitative coding process were conducted by the authors to identify recurring concepts related to AI governance, security, auditability, and assurance. These recurring themes were systematically synthesized into the proposed control domains that constitute the proposed AI audit framework. Since the thematic coding was performed by the authors without independent inter-coder validation, this is acknowledged as a limitation of the study.

4.2. Development of the Control Framework

The control framework developed within this study is designed to mitigate risks encountered throughout the AI system life cycle (Falco et al., 2021). In constructing the framework, the control requirements identified in the literature review were grouped thematically and structured by aggregating controls of a similar nature.

To facilitate a more effective analysis of control items and to increase applicability across diverse organizational contexts, the framework is structured around four core components:

Domain: Represents the primary category to which the control belongs (e.g., documentation, compliance, application security, model security).

Risk Level: Categorizes the priority of the control based on a risk-based assessment approach (NIST, 2023).

Control Title (Subject): Defines the specific control requirement addressed within the scope of the audit.

Description: Details the scope, implementation method, and specific requirements of the control.

This structure ensures that control items address both technical and administrative dimensions, resulting in an integrated model that allows for multidisciplinary evaluation.

4.3. Classification of Control Domains

Following thematic analysis, the developed control framework was classified into five primary control domains to ensure the conceptual integrity of requirements and to establish a systematic approach for audit processes.

It is important to distinguish between the conceptual architecture of the framework and its operational implementation. The four conceptual dimensions define the principal objectives of AI auditing, whereas the five control domains represent the practical organization of audit controls. Consequently, the increase from four conceptual dimensions to five operational domains does not indicate an expansion of the framework but rather a decomposition of governance and data-related requirements into more manageable and auditable control structures. This distinction improves both conceptual clarity and implementation feasibility.

4.3.1. Documentation Controls

Documentation controls are critical for the transparency, traceability, and sustainability of AI systems (Mitchell et al., 2019). This domain addresses requirements such as maintaining an AI component inventory, documenting model and dataset metadata, defining responsibilities, and establishing acceptable use policies. These controls primarily contribute to ensuring accountability during audit processes (Gebru et al., 2021).

4.3.2. Compliance Controls

Compliance controls aim to ensure that AI systems adhere to legal regulations and ethical standards. This includes evaluating the reliability and legal conformity of data sources, requirements for the protection of personal data (e.g., General Data Protection Regulation (EU) 2016/679 (GDPR)/ The Personal Data Protection Law (KVKK)), data usage permissions, and AI impact assessments (Kaminski & Malgieri, 2021). These controls play a vital role in managing data-centric risks.

4.3.3. Access and Authorization Controls

These controls aim to prevent unauthorized access to AI systems and training datasets. Defined measures include access restrictions for training sets, authorization mechanisms, change management processes, and protections against unauthorized interventions, thereby contributing to data integrity and system security (Saltzer & Schroeder, 1975).

4.3.4. Application Security Controls

Application security controls focus on mitigating security risks during the development and operation of AI systems. Elements such as the Secure Software Development Life Cycle (Secure SDLC), input/output validation mechanisms, code review processes, vulnerability management, the use of sandboxes, prompt security, and cache protection are addressed here (OWASP, 2023). This domain is particularly vital for the secure operation of LLM-based systems.

4.3.5. Model Security Controls

These controls are directed toward ensuring the integrity, accuracy, and security of AI models. Evaluated measures include model integrity verification, accuracy testing, performance testing, management of overfitting and underfitting risks, protection methods against adversarial attacks, and secure model deployment (Madry et al., 2019). These are critical for preventing model-specific exploits.

4.4. Risk-Based Classification Approach

Control items are classified into three levels based on a risk-based approach, enabling organizations to prioritize requirements and conduct audits effectively (ISO/IEC 23894, 2023):

Level 1 (High Priority): Mandatory controls encompassing critical security requirements.

Level 2 (Medium Priority): Controls that enhance system security but may vary according to organizational context.

Level 3 (Low Priority / Advanced): Controls requiring higher maturity levels and involving advanced security measures.

4.5. Presentation of the Control Framework

The proposed framework organizes AI audit controls into structured control domains and defines the principal categories that should be considered during audit planning and implementation. Rather than presenting an exhaustive operational catalogue, the framework provides a conceptual structure that can guide future development of detailed audit controls. This systematic arrangement provides a clear hierarchy that facilitates both academic analysis and practical implementation. By presenting the principal control domains and representative control categories, the study provides a structured reference framework that supports practitioners and auditors in evaluating AI systems against key security and governance requirements.

4.6. Alignment with Standards

The proposed framework was developed to remain compliant with international standards in information security and AI governance. Reference was made to frameworks such as ISO/IEC 27001, ISO/IEC 42001, NIST AI Risk Management Framework, and the OWASP Top 10 for LLM Applications. By ensuring conceptual and structural alignment with these standards, the model provides an integrable and applicable structure for both academic research and industry practice (Anderljung et al., 2023).

4.7. AI Assurance Levels

While the operational control domains define **what should be audited**, they do not specify **how extensively these controls should be verified**. Organizations deploy AI systems with substantially different risk profiles, operational environments, and regulatory obligations. Consequently, a uniform audit depth would either overburden low-risk systems or underestimate the risks associated with high-impact AI applications. To address this challenge, the proposed framework introduces Artificial Intelligence Assurance Levels (AIAL), which determine the depth of audit activities according to system criticality and risk exposure. In this way, the framework combines implementation-oriented controls with scalable assurance mechanisms.

Artificial intelligence systems (AI) exhibit significant differences in terms of their intended use, the types of data they process, their decision-making capabilities, and their operational impact. Therefore, evaluating all AI systems with the same assurance requirements is neither technically nor managerially feasible. For example, subjecting an AI system used for corporate document summarization to the same evaluation scope as one used to support health diagnostic processes or influence access to public services can lead to disproportionate results.

This situation necessitates the creation of assurance mechanisms compatible with the risk levels of AI systems. Similarly, the ISO/IEC 15408 standard, used in the security assessment of information technology products, also stipulates that assessment activities should be structured according to different assurance levels. However, due to the data dependency, variability in model behavior, and continuous learning characteristics of AI systems, this approach needs to be reinterpreted in a way specific to AI systems.

In this context, the proposed methodology develops a four-level Artificial Intelligence Assurance Level (AIAL) model to systematically classify the scope and depth of assessment activities. The appropriate assurance level is determined according to the Artificial Intelligence System Risk Level (AISRL) assigned to the audited AI system through the proposed risk assessment process.

4.7.1. Structure of Assurance Levels

In our proposed model, the assurance levels consist of four levels that progressively increase the expected assurance depth for AI systems (Table 2). AISRL represents the inherent risk classification of an AI system. Based on the assigned ASRL, the corresponding Artificial Intelligence Assurance Level (AIAL) determines the extent of audit activities, evidence requirements, and assurance procedures. Each level covers the requirements of the previous level and includes additional assurance activities. This approach ensures that as assurance levels increase, more evidence is collected, more comprehensive technical assessments are carried out, and higher-level governance mechanisms are implemented.

Table 2. AI Assurance Levels

Assurance Level	Assurance Approach	Main Objective
ASRL-1	Basic Assurance	Verification of minimum security and governance controls.
ASRL-2	Structural Assurance	Verification of the systematic application of processes and control mechanisms.
ASRL-3	Advanced Assurance	Assessing resilience against AI - specific threats.
ASRL-4	Continuous Assurance	Continuous monitoring and dynamic assurance throughout the lifecycle.

(Table created by authors)

The fundamental difference between assurance levels lies not in the controls implemented, but in the scope and depth of the assessment activities. Lower levels verify the presence of basic safety and governance controls, while higher levels involve technical safety testing, adverse analyses, and continuous monitoring mechanisms.

This approach, like the Common Criteria assessment logic, is based on the principle that as assurance levels increase, the amount of evidence evaluated and the scope of verification activities should also increase. However, unlike traditional information technology products, the proposed model also includes AI system-specific dimensions such as

data security, model security, prompt security, and the continuity of model behavior within its scope of evaluation (Table 3).

Table 3. Increasing Depth of Assurance by Level

Evaluation Dimension	ASRL-1	ASRL-2	ASRL-3	ASRL-4
Documentation Review	✓	✓	✓	✓
Process Verification		✓	✓	✓
Technical Safety Tests			✓	✓
Adversary Evaluation			✓	✓
Drift Analysis			✓	✓
Continuous Monitoring				✓
Continuous Re-evaluation				✓

(Table created by authors)

Each assurance level includes specific technical and administrative control requirements. These requirements are structured to manage the risks that may arise throughout the lifecycle of AI systems.

Administrative controls aim to ensure that artificial intelligence systems are operated reliably, not only from a technical standpoint but also in terms of governance and organizational processes. Current studies show that a significant portion of the risks arising in artificial intelligence systems stem not from technical vulnerabilities but from inadequate governance, unclear responsibilities, incomplete risk management processes, and insufficient human oversight (Brundage et al., 2020). Therefore, to create reliable artificial intelligence systems, technical security measures must be supported by corporate governance mechanisms. Furthermore, the ISO/IEC 42001 standard recommends defining policies, responsibilities, risk management, and continuous improvement processes throughout the lifecycle of artificial intelligence systems.

The following control areas are evaluated within this scope:

- AI Governance
- Risk Management
- Roles and Responsibilities
- Human Oversight
- Incident Response Processes
- Compliance and Ethics Management

On the other hand, technical controls aim to protect artificial intelligence systems against security threats that may arise in the data, model, and application layers. Unlike traditional information systems, artificial intelligence systems are exposed to AI-specific threats such as data poisoning, model extraction, membership inference, prompt injection, and adversarial attacks (Goodfellow et al., 2015). Therefore, technical security assessments are not limited to classic mechanisms such as access control and network security; they also encompass data security, model security, output security, and operational monitoring activities.

Within this scope, the following technical control areas are evaluated:

- Data Security
- Model Security
- Prompt Security
- Access Control
- Logging and Monitoring
- Drift Detection
- Adverse Resistance

Table 4. Audit Requirements According to Assurance

Control Domain	ASRL-1	ASRL-2	ASRL-3	ASRL-4
AI Policies	✓	✓	✓	✓
Risk Management		✓	✓	✓
Data Security	✓	✓	✓	✓
Model Security		✓	✓	✓
Prompt Security		✓	✓	✓
Adversarial Tests			✓	✓
Drift Monitoring			✓	✓
Continuous Monitoring				✓
Continuous Assurance				✓

(Table is created by the authors)

Since artificial intelligence systems have different risk profiles and usage contexts, it is not appropriate to evaluate all systems with the same audit and assurance requirements. It is shown in Table 4. Therefore, in the proposed methodology, a four-level AI Assurance Level model is defined to systematically differentiate the scope and depth of evaluation activities. Assurance levels determine the scope of the technical and administrative controls expected to be implemented, as well as the depth of the verification activities to be carried out in the evaluation process. As the level increases, assurance activities expand from basic documentation reviews to technical safety tests, adversarial assessments, and continuous monitoring mechanisms. This approach is based on a tiered assurance understanding and aims to increase the intensity of the evaluation proportionally to the risk and criticality level of the system. Table 5 presents the proposed AI Assurance Levels, the corresponding assurance depths for these levels, and the basic control requirements expected to be implemented.

Table 5. AI Assurance Levels, Depth of Assurance, and Control Requirements

Assurance Level	Assurance Objective	Assurance Depth	Administrative Controls	Technical Controls
ASRL-1 Main Assurance	Verification of fundamental security and governance requirements.	Documentation review and basic configuration verification.	AI policies, roles and responsibilities, logging.	Authentication, access control, logging.
ASRL-2 Structural Assurance	Verification of the systematic implementation of security processes	Process reviews and technical verification activities.	Risk management, data governance, human oversight, lifecycle management	Data integrity, model access control, output filtering, secure configuration.
ASRL-3 Advanced Assurance	Assessing resilience against AI-specific threats.	Technical safety tests and adverse evaluations.	Incident response plans, safety testing procedures, compliance assessments	Prompt injection tests, adversarial tests, model extraction, and data poisoning analyses.
ASRL-4 Continuous Assurance	Maintaining safety throughout the lifecycle.	Continuous monitoring, reassessment, and dynamic assurance.	Continuous governance, periodic audits, continuous risk management.	Drift monitoring, behavioral analysis, real-time anomaly detection, continuous adverse monitoring.

(Table is created by the authors)

In the proposed model, assurance levels define not only the control sets to be applied but also the scope and depth of the assessment activities. As the assurance level increases, assurance activities expand from documentation review to technical verification, adversarial assessments, and continuous monitoring mechanisms. This approach is inspired by the tiered assurance concept in the ISO/IEC 15408 standard, but also encompasses assessment areas specific to artificial intelligence systems, such as data security, model behavior, prompt security, and model lifecycle. Thus, a direct relationship is established between the assurance level and the control requirements and assessment depth.

4.7.2. Determining Assurance Levels for Artificial Intelligence

The risk profiles of AI systems vary significantly depending on factors such as the context of use, the nature of the data processed, decision-making authority, and the operational impact of the system. Therefore, it is neither technically nor regulatorily feasible to evaluate all AI systems under the same assurance requirements. Risk-based governance approaches developed in recent years also suggest that AI systems should be evaluated at different levels according to their contextual risks (NIST, 2023; European Parliament and Council, 2024).

In the proposed methodology, assurance levels are determined by quantitatively assessing the risks to which the system is exposed. This approach allows for the application of more comprehensive assurance requirements to high-risk systems while reducing the compliance burden on low-risk systems. Thus, assessment activities are made proportional to risk. The methodology is inspired by the Evaluation Assurance Level (EAL) approach in the ISO/IEC 15408 standard, which envisages a gradual increase in assessment depth according to assurance levels. However, the proposed AI Assurance Levels are redesigned to take into account risk factors specific to AI systems, such as data dependency, model behavior, autonomy level, and operational impact, unlike the security assessment of traditional IT products. In this respect, the framework preserves the assurance logic of the Common Criteria approach while offering a risk-based assessment model suitable for the dynamic and adaptive nature of AI systems. In addition, the systematic assessment and evidence-based verification principles defined in the ISO/IEC 18045 standard have been utilized in structuring the assessment process. The risk dimensions in Table 6 were taken into account in determining the assurance level. The risk criteria defined in NIST AI RMF (2023), ISO/IEC 23894:2023 and the European Union AI Act were used in the selection of dimensions.

Table 6. Core Dimensions of AI System Risk Classification

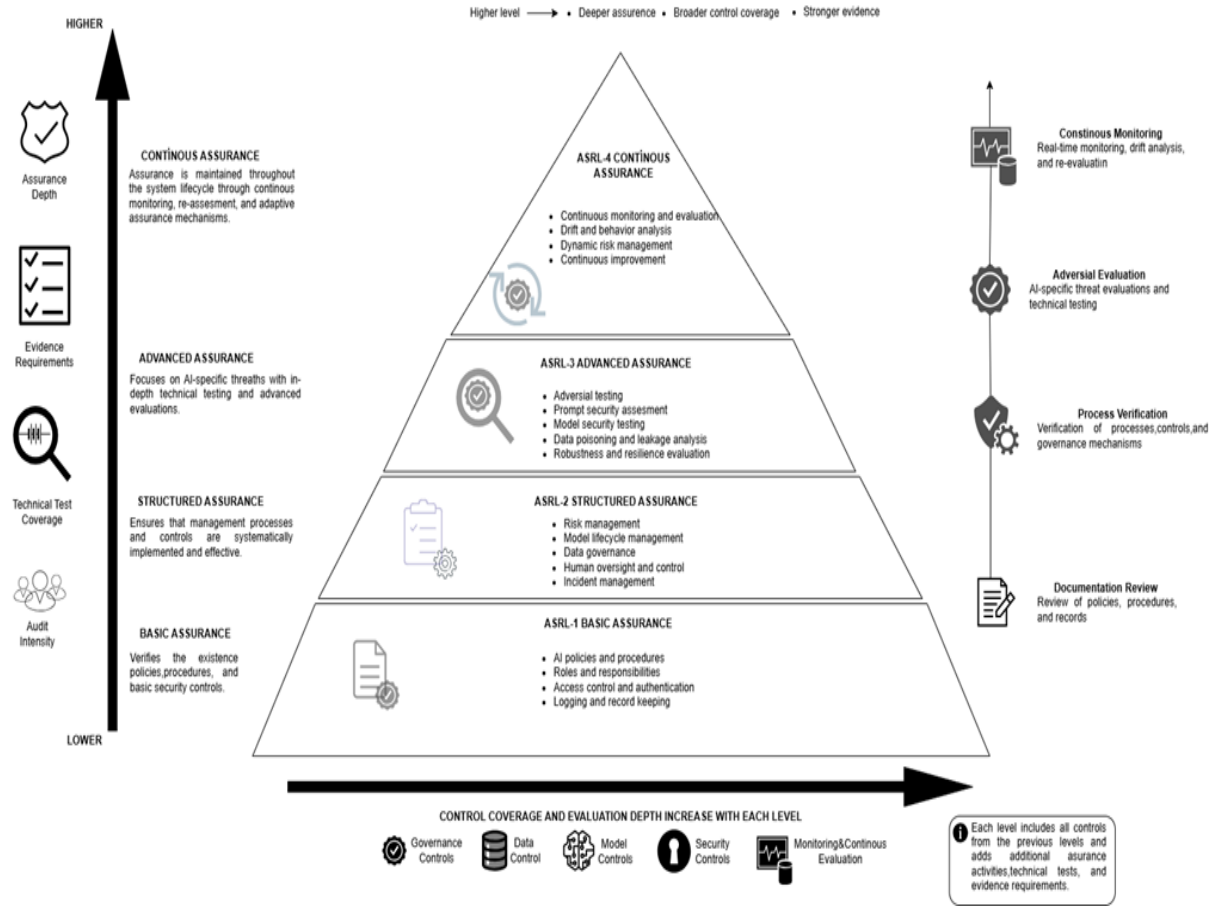
Risk Dimension	Description
Data Sensitivity	The level of confidentiality and criticality of the processed data.
Decision Impact	The impact of AI outputs on individuals or processes.
Level of Autonomy	The capacity to make decisions without human intervention.
Usage Scale	Number of affected users or transactions
External Access	Integration with the internet or third-party systems
Regulatory Impact	Legal or sectoral obligations

(Table is created by the authors)

The risk profiles and usage contexts of artificial intelligence systems vary significantly. Therefore, it is not appropriate to evaluate all systems with the same assurance requirements. In the proposed methodology, a four-level AI Assurance Level model is defined to differentiate evaluation activities according to system criticality and risk level. These levels systematically classify the scope of assurance requirements, the technical and administrative controls to be applied, and the depth of evaluation. As the assurance levels increase, the scope of evaluation expands, technical verification activities deepen, and continuous assurance requirements come to the forefront.

The assurance-oriented structure proposed in this study is illustrated in Figure 1, which presents a hierarchical view of AI audit assurance levels and their corresponding control and evidence requirements. The model reflects the principle that higher assurance levels require greater audit depth, stronger verification activities, and more comprehensive evidence collection. By establishing a systematic relationship between risk exposure, control implementation, and assurance expectations, the framework provides organizations with a practical roadmap for scaling AI audit activities according to the criticality and risk profile of AI systems.

Figure 1. Assurance Pyramid



(Figure is created by the authors)

5. CONCEPTUAL EVALUATION OF THE PROPOSED AI AUDIT FRAMEWORK

5.1. General Findings

Proposed AI audit and assurance framework has been developed as a design-oriented model within the scope of this study provides a comprehensive security and governance structure across diverse control domains. The analysis reveals that risks associated with Artificial Intelligence (AI) systems are multidimensional, necessitating evaluation not only through the lens of model performance but also across the dimensions of data security, ethical compliance, systemic integrity, and operational security (Floridi et al., 2018; Raji et al., 2020). This analytical discussion aims to demonstrate the theoretical consistency and practical applicability of the proposed methodology while acknowledging that future empirical validation remains necessary.

An examination of the thematic distribution of control objectives indicates that risks are concentrated along four primary axes:

- Data-centric risks (e.g., privacy, quality, and provenance)
- Model security risks (e.g., adversarial robustness and integrity)
- Application layer vulnerabilities (e.g., integration and interface security)
- Governance and compliance deficiencies (e.g., accountability and regulatory alignment)

This distribution underscores that auditing AI systems requires a transdisciplinary approach, where administrative and ethical dimensions carry as much weight as technical controls.

5.2. Evaluation by Control Domains

The proposed control domains should not be interpreted as independent security mechanisms but rather as complementary components of an integrated AI auditing architecture. Each domain addresses a distinct category of organizational or technical risk identified throughout the literature review. Together, these domains establish a comprehensive audit structure covering governance processes, regulatory compliance, data protection, software security, and AI model assurance throughout the system lifecycle.

5.2.1. Documentation and Transparency

Analysis of the documentation controls within the framework demonstrates that transparency and traceability are central to the efficacy of the audit process. The presence of controls directed at the systematic recording of model inventories, dataset sources, and hyperparameter configurations directly correlates with the level of system auditability (Brundage et al., 2020).

Deficiencies in documentation not only impede technical verification processes but also pose significant risks regarding accountability and legal compliance. Consequently, transparency mechanisms are identified as a foundational audit requirement, as their absence exacerbates technical, ethical, and juridical risks.

5.2.2. Compliance and Legal Requirements

Compliance-related analyses indicate that data protection, ethical assessment, and legal adherence constitute critical risk areas. The scope of controls regarding the traceability of data sources, usage permissions, and AI Impact Assessments (AIIA) serves as the primary determinant of an organization's regulatory alignment. These findings suggest a mandatory shift toward systematizing compliance processes to address the evolving global regulatory landscape (Jobin et al., 2019).

5.2.3. Access and Authorization Management

Findings regarding access and authorization controls highlight that restricted access to data and model components is a critical security parameter. Regulating access to training datasets and model weights is decisive in preserving both data integrity and model reliability. While these controls align with classical information security paradigms, they must be extended to encompass AI-specific risks, such as unauthorized model extraction or data leakage via inference.

5.2.4. Application Security

Observations in application security reveal that AI systems introduce novel threat vectors that transcend traditional software vulnerabilities. Specifically, in Large Language Model (LLM)-based systems, risks such as prompt injection, output validation failures, and insecure integrations represent significant control gaps (OWASP, 2023). This necessitates the expansion of the Secure Software Development Life Cycle (S-SDLC) to integrate AI-specific security testing, as conventional methodologies prove insufficient for these dynamic architectures.

5.2.5. Model Security

Analyses within the domain of model security confirm that AI models are inherently susceptible to specialized attack modalities. Risks such as data poisoning, model inversion, and membership inference attacks threaten the very core of model integrity and privacy (Papernot et al., 2018). Furthermore, the implementation of controls for model validation, integrity verification, and secure deployment is vital for establishing trust. These findings justify the treatment of model security as a distinct and autonomous control domain within the audit framework.

5.3. Risk-Based Prioritization Analysis

The distribution of control objectives according to risk levels shows a significant weighting toward high-priority controls. This suggests that fundamental security and governance mechanisms are indispensable for the secure operation of AI systems. Conversely, medium and low-priority controls generally encompass advanced security measures applicable as system maturity increases. This highlights the necessity for organizations to adopt a risk-based prioritization approach in their AI auditing cycles (NIST, 2023).

5.4. Comparison with Existing Standards

The proposed framework exhibits substantial alignment with prevailing international standards and frameworks:

- NIST AI Risk Management Framework: Alignment in risk-oriented methodology.
- ISO/IEC 27001: Convergence on foundational information security controls.
- ISO/IEC 42001: Integration of AI governance structures.
- OWASP Top 10 for LLM Applications: Alignment regarding application-level vulnerabilities.

However, this study provides an added contribution by offering more granular control requirements specifically tailored to model security, prompt management, and LLM-specific risks, which are currently treated at a higher level of abstraction in existing standards.

5.5. Synthesis and Conclusion

The findings collectively demonstrate that AI auditing requires a multi-layered architecture. The integration of technical controls, data management, ethical compliance, and governance mechanisms forms the bedrock of an effective audit process. The framework proposed herein addresses this multidimensionality, offering a holistic approach for organizations to manage AI systems in a secure, transparent, and auditable manner. Furthermore, the results indicate that current frameworks lack sufficient detail regarding LLM security and model-specific adversarial threats, marking these as critical focal points for future scholarly and practical research.

The proposed framework provides comprehensive coverage of the principal governance and technical security requirements identified throughout the literature. By integrating organizational controls, application security, data protection, and model assurance within a single audit architecture, the framework offers a structured methodology for conducting AI audits across diverse organizational environments.

Nevertheless, these observations should be interpreted as conceptual outcomes derived from the framework design rather than empirical validation results. Although the proposed methodology appears theoretically consistent and practically implementable, its effectiveness should be confirmed through future case studies, pilot implementations, and comparative evaluations across different industrial sectors.

6. DISCUSSION

The AI audit control framework proposed in this study transcends traditional information security approaches found in current literature by offering a holistic model that encompasses risks unique to artificial intelligence systems (Falco et al., 2021; Raji et al., 2020). In light of the findings, it is evident that the auditing of AI systems cannot be confined to technical controls alone; rather, it necessitates a multidimensional approach that concurrently addresses data, model, application, and governance layers (Floridi et al., 2018; Gasser & Almeida, 2017).

An examination of existing standards reveals that while frameworks such as ISO/IEC 27001 provide general security principles, they remain limited regarding AI-specific risks—particularly model behavior, data dependency, and learning processes (Mökander et al., 2021). Conversely, next-generation standards like the NIST AI Risk Management Framework and ISO/IEC 42001 offer a significant foundation for AI governance but lack detailed, implementation-level control objectives (Schiff et al., 2020).

In this context, the proposed control framework contributes to the literature primarily in the following areas:

- Granularity of LLM Security: Providing detailed security controls specifically tailored for Large Language Models (OWASP, 2023).
- Addressing Emerging Threat Vectors: Handling novel threats such as prompt security, data leakage, and model manipulation (Papernot et al., 2018).
- Integrated Evaluation: Assessing model security and application security within a unified framework.
- Contextualized Risk Management: Adapting a risk-based control approach specifically for the lifecycle of AI systems (Amodei et al., 2016).
- Assurance-Oriented Audit Architecture: This approach allows for the definition of graduated assurance levels and associated evidence requirements for AI system audits by introducing an assurance-based control structure

that is inspired by well-known security evaluation methodologies, especially the Common Criteria framework (ISO/IEC 15408; CCRA,2022).

Nevertheless, this study is subject to certain limitations. First, the proposed framework was developed as a theoretical and conceptual model and has not yet been empirically tested across diverse industry sectors. Furthermore, the feasibility of implementing these control items may vary significantly depending on an organization's technical capacity, available resources, and maturity level (Metcalf et al., 2021).

Another critical point of discussion is the dynamic nature of AI systems. Unlike traditional software systems, AI models can update their behaviors over time in response to shifting data and conditions (Anderson, 2020; Jordan & Mitchell, 2015). This characteristic implies that audit processes must evolve from static evaluations into continuous and adaptive frameworks (Brundage et al., 2020).

Additionally, the work contributes to the emerging domain of AI assurance by delineating a systematic relationship among control requirements, audit evidence, and assurance levels. The current AI governance frameworks mostly emphasize risk detection and management; however, the proposed approach broadens this viewpoint by integrating assurance principles typically used in security assessments. This allows enterprises to identify AI-related hazards and to provide verifiable evidence that necessary safeguards have been deployed and are functioning properly.

Finally, the necessity of integrating ethical and legal dimensions with technical controls must be emphasized. Ensuring principles such as transparency, accountability, and fairness in AI decision-making processes is achievable not solely through technical measures, but through the robust implementation of governance mechanisms (Kroll et al., 2015; Cath, 2018).

From a practical perspective, the proposed framework may assist audit functions in integrating AI systems into risk-based audit planning and assurance activities. By providing a structured approach for identifying critical governance, security, and compliance controls, the proposed framework enables auditors to prioritize audit efforts according to the characteristics and potential risks of different AI systems. It also supports the consistent evaluation of documentation, access control, application security, and model security throughout the AI lifecycle, thereby facilitating more systematic and transparent audit practices. Although the proposed framework has not yet been empirically validated, it offers a practical conceptual foundation that organizations can adapt to strengthen AI governance, improve audit planning, and support future assurance activities. Future empirical studies and organizational case applications may further evaluate and refine the proposed approach.

7. CONCLUSION AND RECOMMENDATIONS

7.1. Conclusion

In this study, a comprehensive and multi-layered control framework for auditing artificial intelligence systems has been developed. The proposed model encompasses five primary domains—documentation, compliance, access control, application security, and model security—and is structured through a risk-based approach (NIST, 2023; Raji et al., 2020).

The results of the study suggests that the following elements are critical for the secure and effective management of AI systems:

- Security and transparency of data sources (Brundage et al., 2020).
- Ensuring model integrity and accuracy (Papernot et al., 2018).
- Implementation of security controls at the application layer (OWASP, 2023).
- Effective management of authorization and access mechanisms (Anderson, 2020).
- Ensuring legal and ethical compliance (Floridi et al., 2018; Jobin et al., 2019).

These findings demonstrate that AI auditing is not merely a technical procedure but an interdisciplinary field requiring a governance and risk management perspective (Gasser & Almeida, 2017; Falco et al., 2021). The synthesis of these critical elements underscores that robust AI auditing must transcend the traditional checklist approach, evolving instead into a dynamic, lifecycle-oriented governance strategy. By integrating data provenance, algorithmic integrity, and rigorous application security with high-level ethical and legal mandates, organizations can move beyond mere technical verification toward a state of verifiable trustworthiness. This interdisciplinary alignment not only fortifies AI systems against emerging adversarial threats but also ensures that the rapid pace of innovation remains anchored in

accountability. Ultimately, the successful operationalization of this multi-layered framework is essential for fostering the sustainable and transparent deployment of artificial intelligence across critical sectors of society.

7.2. Recommendations

Based on the findings, the following recommendations are offered to organizations developing or utilizing AI systems:

- Adoption of a Risk-Based Audit Approach: Controls for AI systems should be prioritized and implemented according to established risk levels (NIST, 2023).
- Establishment of AI Governance Structures: Organizations should clearly define policies, procedures, and responsibilities regarding AI usage (ISO/IEC 42001, 2023; Cath, 2018).
- Enhancement of Data Management and Transparency: Training datasets, data sources, and data processing workflows must be made traceable and auditable (Kroll et al., 2015).
- Prioritization of Model Security: Mechanisms for model validation, integrity verification, and protection against adversarial attacks should be implemented.
- Securing LLM and Generative AI: Prompt security, data loss prevention (DLP), and output monitoring mechanisms should be developed (OWASP, 2023).
- Establishment of Continuous Audit Mechanisms: Given the dynamic nature of AI systems, audit processes should be continuous rather than periodic (Mökander et al., 2021; Falco et al., 2021).

The implementation of these recommendations serves as a vital bridge between theoretical risk frameworks and the operational realities of modern enterprise environments. By transitioning from traditional, static oversight to a dynamic, multi-layered governance model, organizations can effectively mitigate the unique vulnerabilities inherent in machine learning lifecycles. Furthermore, fostering a culture of transparency and continuous verification not only addresses immediate security concerns but also cultivates the institutional trust necessary for the long-term integration of generative AI. Ultimately, the systematic adoption of these controls ensures that as AI capabilities evolve, the mechanisms designed to protect and govern them remain equally resilient, ethical, and aligned with global regulatory trajectories.

7.3. Future Work

It is suggested that the control framework developed within the scope of this study be expanded in the following directions:

- Empirical Testing: Applying and testing the framework through empirical studies across various sectors (e.g., finance, healthcare, public sector).
- Maturity Model Integration: Supporting the control framework with a formal maturity assessment methodology and a comprehensive operational control catalogue to facilitate the practical implementation of the proposed framework and measure organizational progress (Caralli et al., 2010; Metcalf et al., 2021).
- Automation: Integration with automated audit tools and AI-assisted auditing systems.
- Regulatory Alignment: Conducting detailed analyses of the framework's alignment with emerging AI regulations, such as the EU AI Act (Schiff et al., 2020).

Future research in these directions will play a pivotal role in the evolution of the field, transforming theoretical auditing concepts into robust, operationalized methodologies. By empirically validating these controls across high-stakes industries and integrating them with automated diagnostic tools, the academic and professional communities can bridge the gap between high-level ethical principles and granular technical verification. Ultimately, such advancements will contribute to the development of more applicable, scalable, and measurable models that ensure AI systems remain not only technologically superior but also demonstrably secure, transparent, and aligned with the evolving global regulatory landscape.

References

Acemoglu, D., & Restrepo, P. (2019). "Automation and New Tasks: How Technology Displaces and Reinstates Labor." Journal of Economic Perspectives, 33(2), 3–30.

Amodei, D., et al. (2016). "Concrete Problems in AI Safety." arXiv preprint arXiv:1606.06565.

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media & Society, 20(3), 973–989.

Anderljung, M., Barnhart, J., Korinek, A., et al. (2023). Frontier AI Regulation: Managing Emerging Risks to Public Safety. arXiv preprint arXiv:2307.03718.

Anderson, R. (2020). Security Engineering: A Guide to Building Dependable Distributed Systems (3rd ed.). Wiley.

Ashmore, R., et al. (2021). Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges. ACM Computing Surveys.

Barocas, S., Hardt, M., & Narayanan, A. (2023). Fairness and Machine Learning: Limitations and Opportunities. MIT Press.

Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition, 84, 317–331.

Birhane, A., Steed, R., Ojewale, V., Vecchione, B., & Raji, I. D. (2024). AI auditing: The Broken Bus on the Road to AI Accountability. Proceedings of the 2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML).

Brundage, M., et al. (2020). Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. arXiv preprint arXiv:2004.07213 / GovAI Technical Report.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data & Society, 3(1).

Caralli, R. A., et al. (2010). CERT Resilience Management Model (CERT-RMM): A Maturity Model for Managing Operational Resilience. Addison-Wesley Professional.

Cath, C. (2018). "Governing artificial intelligence: ethical, legal and technical opportunities and challenges." Philosophical Transactions of the Royal Society A.

Cobbe, J., Lee, M. S., & Singh, J. (2020). Reviewable Automated Decision-Making: A Framework for Accountability. Computer Law & Security Review.

Committee of Sponsoring Organizations of the Treadway Commission (COSO). (2023). Realize the full potential of artificial intelligence: Applying the COSO framework and principles to help implement and scale artificial intelligence.

Common Criteria Recognition Arrangement. [CCRA], (2022). Arrangement on the recognition of Common Criteria certificates in the field of information technology security. Common Criteria Portal.

Costanza-Chock, S. (2020). Design Justice: Community-Led Practices to Build the Worlds We Need. MIT Press.

European Parliament, & Council of the European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. Official Journal of the European Union, OJ L, 2024/1689.

Falco, G., Shneiderman, B., Badger, J., et al. (2021). Governing AI Safety Through Independent Audits. Nature Machine Intelligence, 3(7), 566–571.

Floridi, L., et al. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds and Machines, 28(4), 689–707.

Gasser, U., & Almeida, V. A. (2017). "A Layered Model for AI Governance." IEEE Internet Computing.

Gebru, T., Morgenstern, J., Vecchione, B., et al. (2021). Datasheets for Datasets. Communications of the ACM, 64(12), 86–92.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. International Conference on Learning Representations (ICLR).

Guidotti, R., et al. (2018). A Survey of Methods for Explaining Black Box Models. ACM Computing Surveys (CSUR), 51(5), 1–42.

Hevner, A. R., et al. (2004). Design Science in Information Systems Research. MIS Quarterly, 28(1), 75-105.

ISO/IEC 15408:2022. Information security, cybersecurity and privacy protection — Evaluation criteria for IT security. International Organization for Standardization.

ISO/IEC 18045:2022. Information security, cybersecurity and privacy protection — Evaluation criteria for IT security — Methodology for IT security evaluation.

ISO/IEC 23894:2023. Information technology — Artificial intelligence — Guidance on risk management. International Organization for Standardization.

ISO/IEC 27001:2022. Information security, cybersecurity and privacy protection — Information security management systems — Requirements.

ISO/IEC 42001:2023. Information technology — Artificial intelligence — Management system. International Organization for Standardization.

Ji, Z., et al. (2023). Survey of Hallucination in Natural Language Generation. ACM Computing Surveys.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399.

Jordan, M. I., & Mitchell, T. M. (2015). "Machine learning: Trends, perspectives, and prospects." Science, 349(6245), 255–260.

Kaminski, M. E., & Malgieri, G. (2021). Algorithmic Impact Assessments under the GDPR: Producing Multi-layered Explanations. International Data Privacy Law.

Kroll, J. A., et al. (2015). "Accountable Algorithms." University of Pennsylvania Law Review, 165, 633.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). "Deep learning." Nature, 521(7553), 436–444.

Leocádio, D., Malheiro, L., & Reis, J. (2024). Artificial Intelligence in Auditing: A Conceptual Framework for Auditing Practices. Administrative Sciences, 14(10), 238

Madry, A., et al. (2019). Towards Deep Learning Models Resistant to Adversarial Attacks. International Conference on Learning Representations (ICLR).

Mehrabi, N., et al. (2021). A Survey on Bias and Fairness in Machine Learning, ACM Computing Surveys (CSUR), Volume 54, Issue 6, Article No.: 115, Pages 1 - 35

Metcalf, J., et al. (2021). "Algorithmic Impact Assessments and Accountability." Communications of the ACM.

Metaxa, D., et al. (2021). Algorithm Auditing: Methods, Ethics, and Best Practices. Foundations and Trends® in Human–Computer Interaction.

Mitchell, M., Wu, S., Zaldivar, A., et al. (2019). Model Cards for Model Reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT), 455–464.

Mökander, J. (2023). Auditing of AI: Legal, Ethical and Technical Approaches. Digital Society, 2(3), 49.

Mökander, J., & Floridi, L. (2021). Ethics-based auditing to promote trustworthy AI. Minds and Machines, 31(2), 323–357.

Mökander, J., et al. (2021). "Ethics-Based Auditing of Automated Decision-Making Systems: Nature, Scope, and Limitations." Science and Engineering Ethics.

NIST (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology.

Nussbaum, M. C. (2009). Creating Capabilities: The Human Development Approach. Harvard University Press.

OWASP (2023). Top 10 for Large Language Model Applications. Open Web Application Security Project.

Papernot, N., et al. (2018). "SoK: Security and Privacy in Machine Learning." IEEE European Symposium on Security and Privacy (EuroS&P).

Raji, I. D., et al. (2020). Closing the AI accountability gap: Defining an internal algorithmic auditing framework. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT), 33–44.

Raji, I. D., et al. (2022). Outsider Oversight: Corporate Responsibility and the Role of External AI Audits. Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society.

Saltzer, J. H., & Schroeder, M. D. (1975). The Protection of Information in Computer Systems. Proceedings of the IEEE.

Sambasivan, N., et al. (2021). “Everyone wants to do the model work, not the data work”: Data Cascades in High-Stakes AI. CHI Conference.

Sandvig, C., et al. (2014). Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. Data and Discrimination: Converting Critical Concerns into Productive Inquiry.

Schiff, D., et al. (2020). "Principles to Practices for Responsible AI: Closing the Gap." arXiv preprint arXiv:2006.04707.

Selbst, A. D., et al. (2019). Fairness and Abstraction in Sociotechnical Systems. Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (FAccT).

Wieringa, M. (2020). What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT).