International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2025: 7.089

Enterprise AI Governance Architecture: Integrating Security, Risk, Compliance, and Human Oversight

Kishan Raj Bellala

Independent Researcher, Austin, Texas, U.S.A
ORCID ID: 0009-0007-2327-0993

Abstract: ***Large Language Models (LLMs) and other generative forms of artificial intelligence are rapidly moving into enterprise environments. Challenges surround ensuring the security of such models, managing risks associated with their use, ensuring that all use meets organizational and external compliance requirements, and ensuring adequate human oversight and accountability of the resulting automated decision making. Work to develop AI governance in the enterprise environment has generally taken a variety of approaches (technical, organizational, and in the form of compliance) and has not formed part of a unified and integrated system. As such, while a large amount of work exists addressing various aspects of the challenges above, considerable opportunity remains to address many of the issues that arise across the entirety of the AI system lifecycle. This paper presents an architecture for an Enterprise AI Governance framework that integrates AI security, risk management, compliance, and human oversight. The proposed architecture presents a unified view for the governance of all forms of AI, across all phases of the AI system lifecycle (use cases, data and prompts, models, deployment and operationalization, access control, monitoring and audit, model updates and retirement). The architecture encompasses identity and access management, data security and protection, AI-specific security controls, risk assessment and management, compliance, explainability, human-in-the-loop decision making, and ongoing monitoring of AI systems in production. A corresponding reference enterprise architecture is presented, detailing how such a governance framework integrates with a cloud-based platform supporting enterprise applications, data stores, cybersecurity operations, and business processes. The proposed architecture is illustrated and evaluated using a number of AI-related scenarios with differing levels of inherent risk. The work ultimately aims to present a unified and sufficiently flexible Enterprise AI Governance Architecture that enables the secure, responsible, accountable, and continuously auditable use of LLMs and other forms of AI in the enterprise, as well as providing a foundation for future work addressing the additional challenges of autonomous systems.***

Keywords: Enterprise AI Governance; Large Language Models; Generative AI; AI Security; AI Risk Management; Regulatory Compliance; Human Oversight; Responsible AI; AI Lifecycle Management; AI Governance Architecture; Explainable AI; Continuous Monitoring; Enterprise Architecture; AI Accountability

1. Introduction

As Large Language Models (LLMs) and generative AI continue to rapidly evolve, enterprises can leverage them to automate and support various aspects of their businesses, such as decision-making, knowledge management, customer interactions, and software development. To make the most out of these applications, enterprises integrate them with their data, cloud, applications, and processes (Hagendorff et al. (2023)). The integration of such systems into organizations, however, poses a number of unique governance challenges that extend far beyond the traditional boundaries of software and information technology governance. Unlike typical applications, LLMs generate output on the fly, process very sensitive information within an organization, interact with external services and can exhibit unpredictable behavior such as “hallucination,” bias, and sensitivity to adversarial input (Autio et al. (2024)). This research explores the currently missing enterprise architecture for secure use of LLMs integrating security, AI risk management, regulatory compliance and human interaction. Current approaches for using LLMs in enterprises typically are based on a collection of technically, securely and organizationally independent solutions for addressing the problems mentioned above. As a result, there are accountability gaps and a multitude of interconnected risks for data leakage, prompt injection, model misuse, hallucination, bias and unreliable output. Furthermore, deciding on the right amount of human interaction in AI-driven processes is difficult: too much human interaction destroys the automated processing benefit; too little human interaction leads to excessive risk of errors and other negative consequences. For this reason, this research proposes implementing monitoring, validation and human interaction on the basis of a risk-based governance.

1.1 Motivation

This research is motivated by the necessity to move AI governance in enterprises from a collection of single rules and measures to a governance architecture that integrates all aspects of enterprise governance, such as AI applications, identity & access management, data governance, cybersecurity & incident management, compliance, risk management, applications and humans. The architecture can be used to establish AI risks and put the appropriate measures for security and compliance in place. The ability to hold accountable individuals or processes for AI-related decisions can help to mitigate risks. Furthermore, the system can be used to monitor and audit the AI during the system’s lifecycle (Birhane et al. (2023)). Thus, this research will develop an architecture that outlines a set of principles for enterprise AI governance, while also demonstrating how governance controls can be technically and operationally integrated into an organization’s existing AI workflows. Governance in this research is considered a capability that exists throughout the AI system lifecycle rather than as a single approval event (OWASP (2024).

1.2 Research Gap

Current approaches to responsible AI, risk management, cybersecurity, privacy and compliance, devised by

researchers and companies alike, all have their strengths and focus on individual governance aspects. However, most of them are designed as stand-alone principles or architectures and lack integration into a comprehensive enterprise architecture for LLM-based applications. Moreover, currently there is no architecture that at the same time deals with security, risk management, compliance, human interaction, governance throughout the entire LLM application life cycle and continuous monitoring and that is integrated into the current enterprise applications and business processes. In order to close these gaps this research proposes an integrated enterprise AI-governance architecture that, in a unified manner, deals with all aspects of governance for responsible AI. Architecture supports all phases of the LLM application life cycle and allows for risk-based decisions and adequate human interaction depending on the characteristics of the application and the potential impact it may have (OWASP (2024)) (European Parliament & Council (2024)).

This study addresses four primary research questions. The first RQ addresses security, risk, compliance and human oversight challenges when deploying AI/LLM-driven systems within enterprise settings. The second RQ examines how all of these previously mentioned aspects can be integrated into a single, unified enterprise AI governance framework or architecture. The third RQ explores how individual governance controls can be successfully applied throughout the entirety of the LLM development lifecycle from initial use case identification to eventual system retirement. Finally, the fourth RQ aims to assess how effectively proposed enterprise AI governance architecture can meet the required governance for a variety of different AI applications and use cases with different associated risk levels. This paper makes six key contributions. First, it outlines a unified enterprise AI governance framework and corresponding architecture integrating aspects of security, risk management, compliance and human oversight. Second, it presents a lifecycle-based approach to AI/LLM governance covering all stages from initial use case identification via risk assessment, model selection and validation, through deployment, monitoring, updating and eventual system retirement. Third, it provides an integrated set of governance and control controls mapped against the enterprise AI/LLM lifecycle. These controls encompass identity and access management, data and prompt governance, model governance, security controls, risk assessment, compliance monitoring and audit, human oversight and finally auditability. Fourth, it outlines a risk-based human oversight model that seeks a balance between full automation and adequate human-based accountability by varying the degree of intervention on the basis of use case impact. Fifth, it presents an accompanying enterprise reference architecture for AI/LLM that details integration with applications, data stores, cloud IaaS, enterprise cybersecurity functions and associated business processes. Sixth, its effectiveness is evaluated via scenario analysis applied to a variety of differing AI applications and use cases with differing associated risk.

1.3 Related Work

AI governance, risk management, LLM security, regulatory compliance, and responsible AI have already been addressed from different perspectives by various frameworks, standards, and approaches for enterprise automation. The NIST Artificial Intelligence Risk Management Framework (AI RMF) includes the four functions Govern, Map, Measure, and Manage. The NIST Generative AI Profile extends the NIST AI RMF to the specific risks of generative AI and LLM-based systems. Organizational approaches for the establishment and continuous improvement of Artificial Intelligence Management Systems (AIMS) are provided by the new standard ISO/IEC 42001:2023. Organizational activities and processes can be integrated with AI risk management by using the new standard ISO/IEC 23894:2023. Security risks of LLM applications are described by the OWASP GenAI LLM Top 10, e.g. prompt injection, sensitive information disclosure, data and model poisoning, insecure output handling, supply-chain vulnerabilities, and excessive agency of LLMs (NIST (2023)) (ISO/IEC 42001 (2023)). Special attention has to be paid to these risks if LLM applications are integrated with enterprise data, APIs, applications, and business processes. Regulatory approaches for AI also have to be considered. The European Union Artificial Intelligence Act adopts a risk-based approach and lays down requirements for applicable AI systems covering issues of risk management, transparency, documentation and record-keeping, human oversight, robustness, and cybersecurity. Furthermore, there are responsible AI approaches such as reliability, safety, security, accountability, transparency, explainability, privacy, and fairness of AI decisions. However, human oversight of LLMs can lead to inaccurate, biased, or even inappropriate output. Thus, excessive human intervention would destroy the major benefit of automation, i.e. increased efficiency and better scalability. Therefore, the level of human involvement has to be proportional to the potential impact and risk of the AI application. The existing approaches provide a strong foundation for integrated approaches connecting governance requirements with corresponding applications, data, identity and access management, cybersecurity, monitoring, business processes, and lifecycle management of AI applications (Truhn et al. (2023)).

2. Proposed Enterprise AI Governance Architecture

In this research an Enterprise AI Governance Architecture is proposed that enables the secure and responsible application of LLM-based applications within enterprises. The proposed Enterprise AI Governance Architecture integrates security, AI risk management, compliance, human oversight, application lifecycle management and monitoring into a single governance framework for enterprise applications (National Institute of Standards and Technology [NIST], 2023; International Organization for Standardization & International Electrotechnical Commission [ISO/IEC], 2023a, 2023b) (Bellala, 2025). Instead of treating these different aspects of AI governance as separate organizational or technical topics, the proposed Enterprise AI Governance Architecture maps them to applications, data platforms, identities and access management, cybersecurity and

monitoring & logging as well as business processes and functions (NIST, 2023; ISO/IEC, 2023a). The governance architecture follows a risk-based approach, i.e. the necessary governance and human oversight can be determined based on the characteristics, sensitivity and potential impact of the specific AI application use case (NIST, 2023; ISO/IEC, 2023b).

2.1 Architecture Objectives

This architecture aims to establish consistent governance in organizations that deploy different AI / LLM-based applications within their enterprise. This architecture aims to fulfill five main goals: (1) protect data and AI applications from security threats and unauthorized access; (2) identify, assess, and manage risks related to AI throughout the system life cycle; (3) support and enable compliance with regulations and the organization's policies through documentation, monitoring and audit; (4) provide risk-based human oversight for AI-generated output and decisions; and (5) monitor the AI applications after they are deployed in order to detect changes in performance, security, risk, and compliance (NIST, 2023; Autio et al., 2024; ISO/IEC, 2023a, 2023b).

2.2 Governance Architecture Layers

This new architecture includes several interconnected governance layers. The Enterprise Application and Business Layer represents the business applications and processes that utilize AI in their activities. The AI Application Layer contains LLM-based applications, AI-agents and AI-enabled services. The Data and Prompt Governance Layer manages the data and prompts that are used to interact with AI and includes mechanisms for data classification, privacy, validation and protection. The Model Governance Layer is responsible for model management, including model selection, evaluation, validation and approval as well as model versioning and retirement. The Security and Access Control Layer manages identity and access to applications and services and includes authentication and authorization as well as encryption, threat detection and monitoring. The Risk and Compliance Layer classifies risk and checks for compliance with local and international regulations and conducts audits to verify governance. The Human Oversight and Monitoring Layer provides human reviewers to verify outputs of applications as well as to monitor, log and provide feedback to all layers of the architecture. All the layers are inter-connected, so that a decision in one layer can have influence on the control in another layer. For example, a high-risk classification can lead to more strict access control, additional validation steps, mandatory human approval and increased monitoring (NIST, 2023; Autio et al., 2024; ISO/IEC, 2023a, 2023b; OWASP, 2025; European Parliament & Council of the European Union, 2024).

2.3 Lifecycle-Based Governance

Architecture implements Governance across the complete AI Lifecycles as opposed to Approval in isolation. AI use-cases are agreed for their Business Value with agreed Benefits. Use-cases are risk classified and assessed considering factors such as the sensitivity of the data used, the potential impact of the AI decisions made, any relevant regulatory requirements and the level of human oversight AI will require. Candidate models are evaluated on their Performance, Security, Reliability, Privacy and Use which defines the governing criteria for the approved AI application(s). The AI application(s) are deployed with adequate Identity, Access, Security and Compliance controls. AI applications(s) are continuously monitored, logged and supported by human oversight and managed through incident management to detect and respond to Security Events, Performance Degrade, Inappropriate Output(s) and increased risk. The system can be re-assessed and revalidated as required before resuming Operation. Where an AI application is no longer required or does not conform to Governance criteria it will be formally retired through controlled processes (NIST, 2023; Autio et al., 2024; ISO/IEC, 2023a, 2023b).

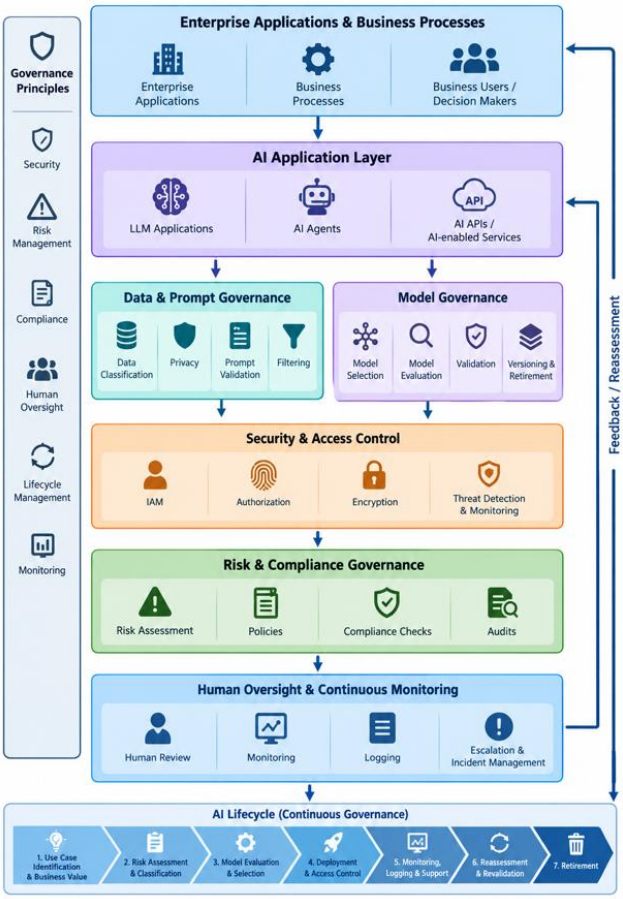

Figure 1: Proposed Enterprise AI Governance Architecture

2.4 Risk-Based Governance and Human Oversight

One of the key elements of the proposed architecture is the use of risk levels to govern appropriate levels of control. Automated control and monitoring may be sufficient for low-risk applications; in addition to automated control and monitoring, medium-risk applications can benefit from occasional human validation. In contrast, high-risk applications must have human approval for all AI-generated output that could result in significant consequences before that output is used for decisions or actions of that nature (NIST, 2023; Autio et al., 2024; European Parliament & Council of the European Union, 2024). This architecture strikes a balance between allowing extensive automation

across an enterprise and ensuring appropriate accountability for use of AI. Risk classification can further determine permissions, data restrictions, model validation, monitoring frequency, audit requirements, and incident response and escalation (NIST, 2023; ISO/IEC, 2023b; OWASP, 2025).

Table 1: Proposed Enterprise AI Governance Architecture Layers

Architecture Layer	Main Purpose	Key Controls
Enterprise Application & Business Layer	Connecting AI with business processes	Business rules, workflows, process ownership
AI Application Layer	Manage LLM-based applications and AI agents	AI services, agents, APIs, application controls
Data & Prompt Governance	Protect information used by AI	Data classification, privacy, prompt validation, filtering
Model Governance	Control model selection and lifecycle	Model evaluation, validation, approval, versioning
Security & Access Control	Protect AI systems and resources	IAM, authentication, authorization, encryption, threat detection
Risk & Compliance	Manage AI risks and regulatory requirements	Risk assessment, policy enforcement, compliance checks, audits
Human Oversight & Monitoring	Maintain accountability and continuous control	Human review, escalation, logging, monitoring, incident management

3. AI Lifecycle and Governance Workflow

The Enterprise AI Governance Architecture is applied throughout the complete lifecycle of an AI/LLM application, from use-case identification in the business to the retirement of a system. The lifecycle of an AI/LLM application consists of use-case identification, including business purpose and benefits, as well as definition of users, required data and necessary AI functionality (Koniakou, 2023; Tabassi, 2023). The use case is then analyzed for risks, such as sensitive data, major business impact, and special regulatory requirements, as well as the degree of autonomy of the AI application and possible security threats (International Organization for Standardization & International Electrotechnical Commission, 2023; Tabassi, 2023). Depending on the identified risk, an appropriate model is selected and checked for performance, reliability, security, privacy and for appropriate application. In addition to the test of the AI application prior to its deployment, validation and approval by the respective governance is also required (Autio et al., 2024; Tabassi, 2023). During the operation of the approved AI application, identity and access control as well as data protection, monitoring, logging and business-process control are deployed. Continuously, all security incidents, incorrect outputs, performance degradation as well as changes in risks and non-compliance are monitored. Substantial changes to the model, the used data, the application itself as well as the business environment require analysis and revalidation by governance prior to continued use (NIST, 2018; Tabassi, 2023). The AI application is retired in a controlled manner, including withdrawal of access, storing of required audit trails and proper handling of data and models, if the AI application is no longer required, does not comply with the agreed-upon governance or has reached the end of its lifecycle (Tabassi, 2023; NIST, 2023).

Table 2: AI Lifecycle and Governance Controls

Lifecycle Stage	Main Activities	Key Governance Controls
Use-Case Identification	Define purpose, users, data, and business process	Business approval, ownership, initial screening
Risk Assessment	Identify and classify AI risks	Risk scoring, impact assessment, regulatory assessment
Model Selection	Select suitable LLM/model	Model evaluation, security and privacy assessment
Validation	Test AI application before deployment	Performance testing, security testing, output validation, compliance review
Deployment	Release approved AI application	IAM, authorization, data protection, logging
Monitoring	Monitor system during operation	Runtime monitoring, audit logs, incident detection, human oversight
Reassessment / Update	Review changes in model, data, or risk	Revalidation, risk reassessment, approval
Retirement	Safely discontinue the AI system	Access removal, data handling, audit retention

4. Integrated Security, Risk, Compliance, and Human Oversight Framework

This framework integrates a number of aspects for the governance of Enterprise AI applications, namely AI/LLM security, prompt security, data protection, risk management, regulatory compliance, human oversight and responsible AI (Tabassi, 2023; Autio et al., 2024; International Organization for Standardization & International Electrotechnical Commission [ISO/IEC], 2023a, 2023b). First, AI/LLM security deals with the security of models, applications, APIs and connected systems of an organization. Prompt security applies filters, checks validation and police prompts to prevent malicious, unwanted, off target or insecure prompts to be processed by an AI model or LLMAI application. Data protection deals with the sensitive data of an organization throughout the AI application life cycle (OWASP, 2025; Autio et al., 2024). It classifies, accesses, processes and protects it as required. Risk classification and risk scoring in this framework define the risk of an AI application or use case. It assesses various factors including data sensitivity, business impact, decision criticality, level of automation and potential consequences of incorrect output (National Institute of Standards and Technology [NIST], 2020; Autio et al., 2024). The outcome of the risk assessment influences the extent of security measures, validation, monitoring and the degree of required compliance with regulations of AI application as well as the involvement of humans in the AI application life cycle. The framework's mechanisms of

regulatory compliance and auditability are based on the enforcement of policies and guidelines, documentation, logging, transparency and traceability of AI decisions (ISO/IEC, 2023b; Tabassi, 2023; Autio et al., 2024). Compliance with requirements and evidence of AI-related governance is monitored and kept throughout the entire AI application life cycle. Humans are involved in AI decision-making either in a human-in-the-loop or human-on-the-loop manner. The first approach requires humans to review and approve output generated by AI applications prior to taking actions based on output (Tabassi, 2023; ISO/IEC, 2023b). The latter enables continuous automated operation while humans monitor the system on an ongoing basis and intervene if required by predefined circumstances or by detection of risk. Therefore, the framework also enables risk-based human approval, i.e. in cases of high impact or high risk of AI decisions and output, a higher degree of human involvement and approval is required than for less risky AI applications that can run automatically to a greater extent. Finally, the framework is also imbued with principles of responsible AI, namely explainability, transparency, prevention of bias and unfairness, reliability and accountability of AI-based decision-making (NIST, 2023; OECD, 2019).

5. Enterprise Case Study and Implementation

To illustrate how the proposed Enterprise AI Governance Architecture can be implemented in a financial enterprise, we consider a representative financial enterprise that comprises multiple business functions such as customer service, loan processing, fraud management, compliance, and financial operations. To support the employees of this financial enterprise, the enterprise deploys an enterprise-level LLM application to support loan application analysis and decision making (Tabassi, 2023; ISO/IEC, 2023a). The AI application, which supports loan application analysis and decision making, is a high-risk AI use case that processes sensitive financial and personal information of customers and makes high-impact financial decisions therefrom. Thus, this case requires strong governance controls and mandatory human oversight (ISO/IEC, 2023b; European Parliament & Council of the European Union, 2024). The proposed architecture is implemented by integrating the AI application with the financial enterprise's existing applications, data platforms, identity and access management systems, cybersecurity monitoring systems, compliance systems, and business workflows (ISO/IEC, 2023a). Before deployment of the use case, the proposed use case is registered with a business owner and undergoes risk assessment based on several factors including data sensitivity, financial impact, regulatory requirements, model reliability, level of automation, and potential security threats (ISO/IEC, 2023b; Tabassi, 2023; Autio et al., 2024). Moreover, the selected LLM is evaluated against several important criteria including accuracy, reliability, security, privacy, bias, and responsibility, to determine whether the LLM is suitable for financial AI-powered decision making (Autio et al., 2024; Tabassi, 2023). The proposed architecture also implements important data and prompt governance controls to restrict access to sensitive customer information, to prevent unauthorized data disclosure, to validate prompts, and to prevent prompt injection attacks that can cause AI applications to behave inappropriately (OWASP, 2025; NIST, 2020). Identity and access management is also an important aspect of the proposed architecture to restrict access to the AI system to only authorized employees and applications. Logging and auditing are also very important to keep records of all user interactions and model outputs, as well as approvers and other significant events. Importantly, the proposed AI system provides AI-generated analysis to a loan officer instead of making automatic loan approval or rejection decisions (ISO/IEC, 2023a; European Parliament & Council of the European Union, 2024). Thus, the proposed Enterprise AI Governance Architecture follows a human-in-the-loop governance model in which a qualified human loan officer reviews the analysis generated by the AI system and makes the final decision therefrom (Rao et al., 2026; Tabassi, 2023). At the same time, the proposed architecture also implements important human-on-the-loop monitoring functions to continuously monitor and observe system performance, security-related incidents, unusual behaviors, and changes in risk. If the system detects a significant security-related incident, unreliable model behavior, unexpected output, or change in relevant regulatory or business requirements, the system can trigger human intervention, risk re-assessment, and model re-validation as appropriate. Importantly, several roles are distributed among various employees in the financial enterprise (Tabassi, 2023; European Parliament & Council of the European Union, 2024).

The business owner of the proposed use case is responsible for the proposed use case, the AI/model governance team is responsible for model validation and governance and model change management throughout the AI application's lifecycle. The security team is responsible for identity and access control, threat detection, and incident response. The risk and compliance teams are responsible for AI-related risk and compliance, the data owners are responsible for classifying, ensuring quality, privacy, and appropriate access to relevant data, and the loan officers or reviewers are responsible for human oversight of high-impact decisions. Importantly, the proposed architecture follows an important sequence of steps from use-case registration to risk assessment, model evaluation, security and compliance validation, governance approval, controlled deployment, continuous monitoring and human review, reassessment and re-validation, to retirement. For example, the monitoring function can prevent high-impact use of the model until the model and associated risks have been re-assessed (Tabassi, 2023; Rao et al., 2026). The Enterprise AI Governance Architecture can thus be implemented on top of a financial enterprise's existing enterprise systems without the need to implement a completely separate AI governance environment (ISO/IEC, 2023a; Koniakou, 2023). Importantly, the proposed security, risk, compliance, responsible AI, and human oversight mechanisms function together to facilitate efficient and responsible use of AI throughout the AI application's complete lifecycle (Tabassi, 2023; Autio et al., 2024; Koniakou, 2023).

6. Evaluation and Comparative Analysis

To determine whether our proposed Enterprise AI Governance Architecture adequately integrates security, risk

management, compliance, monitoring, and human oversight of AI applications within an organization, we apply a scenario-based evaluation using five categories of metrics, namely: security, risk, compliance, monitoring, and human oversight (Tabassi, 2023; Autio et al., 2024). For security, we assess the percentage of security-related threats covered by the appropriate governance controls, number of instances of unauthorized access that are prevented, prompt-injection detection, sensitive data protection, and number of security incidents that are detected (Autio et al., 2024; Open Worldwide Application Security Project, 2024). For risk, we assess the accuracy and consistency of risk classification, number of AI-related risks identified, effectiveness of controls based on risk, and number of high-risk use cases that receive appropriately enhanced governance (ISO/IEC, 2023b; Tabassi, 2023). For compliance, we assess the percentage of appropriate compliance controls, completeness of audit logs, traceability of AI's decision-making and activities, availability of appropriate governance-related documentation, and ability to provide evidence of compliance during an audit. For monitoring, we assess the extent of coverage of deployed AI applications, detection of abnormal or unsafe behavior by models, response time for security incidents, frequency of risk re-evaluation, and detection of model degradation or decreased performance. For human oversight, we assess the percentage of high-risk decisions that are reviewed by a human, percentage of high-consequence actions that are approved by a human, effectiveness of escalation, and rate at which human reviewers can identify and/or reject AI output as inappropriate. The objective is to determine whether the implemented governance controls function correctly and not whether there are policies in place.

We then use three enterprise scenarios to evaluate the architecture. The lowest risk scenario involves summarizing internal documents (e.g. meeting notes) that have been approved for processing by the system (Autio et al., 2024; Tabassi, 2023). The greatest amount of automation would be expected with the lowest amount of required governance. As risk increases to customer-centric scenarios such as customer service chat responses, required data protection, output validation, monitoring and occasional human review are added to the typical amount of automation and governance of the lowest risk scenario (ISO/IEC, 2023b; Tabassi, 2023). The highest risk scenario described in Section 6, a banking loan decision-support application, would have the strongest security as well as require the most extensive model validation, regulatory assessment and continuous monitoring with mandatory human review prior to finalizing a decision of extreme consequence to the customer. We anticipate that architecture would increase the amount of required governance as the potential impact and risk of an AI application increases (ISO/IEC, 2023a; European Parliament & Council of the European Union, 2024).

This architecture can be contrasted with other established approaches to AI governance including the NIST AI Risk Management Framework, ISO/IEC 42001 for an organizational AI management-system, ISO/IEC 23894 which gives a general view on AI risk management, the OWASP guidance for securing LLM applications and the emerging risk-based regulatory approaches such as the EU AI Act which sets out requirements around areas of risk, documentation, transparency, human control, and cybersecurity (Tabassi, 2023; International Organization for Standardization & International Electrotechnical Commission, 2023a, 2023b; Open Worldwide Application Security Project, 2024; European Parliament & Council of the European Union, 2024). As opposed to individual governance dimensions being managed in other architectures, the proposed framework delivers continuous lifecycle governance and maps out the relevant AI risks to corresponding security measures, compliance requirements, levels of monitoring and degree of human control (Tabassi, 2023; Autio et al., 2024). A scenario-based evaluation of the proposed framework's architecture delivers evidence of the value of the governance being constantly applied across all AI risk levels whilst achieving the optimal balance between automation and accountability. However, as with all architecture evaluations, this is a conceptual and scenario-based exercise that does not contain any production measurable data. Therefore, as opposed to describing improved security and organizational performance in quantitative terms, the results only demonstrate the proposed framework's coverage and applicability. Empirical evaluation of this framework in real enterprise AI deployments to quantify and identify improvements to security and organizational performance in terms of number and types of incidents, system audit trails, system monitoring data, and human reviews and their associated cost is required (Tabassi, 2023; Rao et al., 2026).

7. Discussion and Limitations

An evaluation was conducted of the proposed Enterprise AI Governance Architecture. The main value of this architecture is that it enables the governance of enterprise AI applications by linking their required governance with their risk level and their current AI lifecycle stage (Tabassi, 2023; International Organization for Standardization & International Electrotechnical Commission, 2023b). As a result, AI applications with a low risk profile can be automated to a greater extent than applications with a high impact. In addition, AI applications can go through different stages in their AI lifecycle, such as the development, test, pre-production and production stages, and are governed differently in each of these stages (Autio et al., 2024; Tabassi, 2023). The proposed architecture assigns clear roles and responsibilities to the various stakeholders involved in the AI applications, including the business owners, the AI Governance teams, the Security teams, the Risk & Compliance teams, the Data Owners, and the AI application users & operators. These roles and responsibilities can be mapped to the existing organizational structures that are used by enterprises to manage their IT systems and data. The proposed architecture can leverage existing security, identity, data governance, compliance, and monitoring functions that enterprises already have in place (International Organization for Standardization & International Electrotechnical Commission, 2023a; Tabassi, 2023). However, as with any form of good AI Governance, there are also several trade-offs involved. The more rigorous the AI Governance, the larger the cost of implementation, the higher the processing overhead, the longer the time that is required to obtain approval, and the more complex the AI applications become.

In addition, excessive human interaction with the AI applications can destroy their automated nature and reduce their scalability. Also, logging and monitoring of all interactions with the AI applications creates large amounts of data that need to be stored and managed, which can create large security and privacy risks (National Institute of Standards and Technology, 2020; Rao et al., 2026). In addition, it is difficult to assess the risks that are associated with an AI application. The risks of an AI application change as the model, data, users, regulations, and business processes of the enterprise evolve (International Organization for Standardization & International Electrotechnical Commission, 2023b; Autio et al., 2024). As a result, the proposed Enterprise AI Governance Architecture has several limitations. First, the evaluation that was conducted was based on scenarios that are representative of the various use cases that were discussed, but it has not been empirically validated using data from an enterprise that is running the architecture in production (Tabassi, 2023; Rao et al., 2026). Second, the risk scoring and classification that are used in the architecture are not objective but are based on the organizational judgment of the various stakeholders involved. As a result, there is a chance that the scoring and classification are not consistent (International Organization for Standardization & International Electrotechnical Commission, 2023b). Third, human oversight of the AI applications is not sufficient to remove all risks that are associated with the AI applications, since human reviewers can also make mistakes, and they can also overlook certain outputs of the AI applications or become too dependent on the AI applications and start acting irrationally. Fourth, the policies and controls of the architecture need to be updated on a continuous basis as the foundation models, AI agents, security threats, and regulations change. Finally, the implementation of the architecture is more difficult in enterprises that have legacy IT systems, fragmented data environments, limited expertise in AI Governance, and limited monitoring and logging capabilities. In summary, the proposed Enterprise AI Governance Architecture is a flexible governance framework that can be used to support the various requirements for governance of AI applications in an enterprise. However, the framework is not a universal solution, and its effectiveness depends on a number of factors, including the adequate design of organizational policies, clearly defined roles and responsibilities, suitable technical controls, and ongoing evaluation of the AI applications throughout their entire AI lifecycle (Tabassi, 2023; International Organization for Standardization & International Electrotechnical Commission, 2023a; Koniakou, 2023).

8. Future Work

This work will be extended by future research activities in order to cover agentic AI and autonomous enterprise systems. These types of systems plan tasks, interact with various applications, use external tools, and decide and execute actions with only limited human interaction (Autio et al., 2024; Open Worldwide Application Security Project, 2024). Unlike conventional LLM-based applications that only generate content or even provide recommendations, agentic systems are able to execute multiple steps of action. Thus, additional risks need to be considered, for example excessive autonomy, actions that are not authorized, tool misuse, privilege escalation, and other unintended effects on the system (Open Worldwide Application Security Project, 2024; Autio et al., 2024). Adaptive governance mechanisms that adapt security measures, monitoring, risk classification, and human interaction with the AI system in an autonomous manner are needed in future versions of the proposed Enterprise AI Governance Architecture. In addition to these aspects, future research has to investigate continuous risk assessment (Tabassi, 2023; Rao et al., 2026). Here, the governance system has to monitor changes of a model, the used data, user interactions, changed system permissions, and other business-related factors during runtime. In this context, policy-driven autonomous governance is another important direction for future research. Here, autonomous AI systems are restricted by organizational policies. The policies can determine whether high-impact actions of an AI system need to be approved by a human before they can be executed (ISO/IEC, 2023b; Rao et al., 2026) (Autio et al., 2024; Open Worldwide Application Security Project, 2024). Moreover, policies can also determine when an AI system has to stop its actions due to detected risks. Future research has to investigate also the challenges of governance for multi-agent systems. In these systems, multiple AI agents interact with each other in order to support their interaction with enterprise applications. This increases the challenges of accountability, tracking of data flows, authorization, and assignment of responsibility (International Organization for Standardization & International Electrotechnical Commission, 2023a; European Parliament & Council of the European Union, 2024). Empirical evaluation of the proposed architecture for Enterprise AI Governance using real-world enterprise applications could assess the effectiveness of the proposed architecture using security incidents, compliance-related findings, performance of monitoring, results of human reviews, and costs of operation. The architecture can be extended also in order to support emerging regulations and automate compliance check for different geographies and countries. Finally, future research has to investigate how aspects of explainability, fairness, privacy-preserving techniques, and human-AI interaction can be integrated into adaptive AI governance without negatively impacting efficiency of AI systems (Tabassi, 2023; OECD, 2019; Autio et al., 2024).

9. Conclusion

The rapid pace of adoption of enterprise AI and LLM-based applications presents a host of challenges to organizations to address issues of security, AI risk, regulatory compliance, responsible AI, accountability and human oversight (Autio et al., 2024; Tabassi, 2023). While there are existing frameworks and standards for dealing with individual aspects of these challenges, there is a real need for organizations for an integrated framework that can map out the requirements of governance to the AI applications, data, security, processes and people within an organization (Koniakou, 2023; International Organization for Standardization & International Electrotechnical Commission, 2023a, 2023b). The proposed research presents an Enterprise AI Governance Architecture which integrates security, risk management, compliance, human oversight, responsible AI and lifecycle governance. The governance

controls of this architecture are applied throughout the AI lifecycle from use-case identification and risk assessment to model selection and validation, deployment, monitoring, reassessment and retirement (Tabassi, 2023; Autio et al., 2024). A risk-based approach is taken to increase the level of governance and human involvement as the potential impact and risk of application increases. Automation can therefore be increased for lower risk applications while higher risk decisions are subject to greater accountability. A banking case study and scenario-based evaluation is presented to illustrate how the proposed architecture can be applied to implement enterprise AI systems and the mechanisms of governance within the architecture can support issues of security, auditability, monitoring and human decision-making (International Organization for Standardization & International Electrotechnical Commission, 2023b; Tabassi, 2023). The significance of the proposed work is that it provides an enterprise-level integration layer for applying current AI governance, security, risk and compliance practices. The evaluation is conceptual in nature and therefore requires empirical validation. However, it does provide a potential foundation for organizations who are seeking to implement and govern AI systems of increasingly complex nature. As AI applications within organizations evolve towards more agentic and autonomous forms of AI, the architecture provides a potential foundation for the adaptive governance mechanisms that are required to continually assess and manage risk and therefore adjust the controls to ensure that the organization's objectives are met (Autio et al., 2024; Rao et al., 2026). The proposed work therefore provides practical and extensible architecture to support secure, accountable, compliant and risk-aware enterprise AI adoption.

References

[1] Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., & Roberts, K. (2024). Artificial intelligence risk management framework: Generative artificial intelligence profile (NIST AI 600-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.600-1

[2] Birhane, A., Kasirzadeh, A., Leslie, D., & Wachter, S. (2023). Science in the age of large language models. Nature Reviews Physics, 5, 277–280. https://doi.org/10.1038/s42254-023-00581-4

[3] 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). Official Journal of the European Union, L 1689. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

[4] Hagendorff, T., Fabi, S., & Kosinski, M. (2023). Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. Nature Computational Science, 3, 833–838. https://doi.org/10.1038/s43588-023-00527-x

[5] International Organization for Standardization, & International Electrotechnical Commission. (2023). Information technology—Artificial intelligence—Management system (ISO/IEC Standard No. 42001:2023). International Organization for Standardization.

[6] National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1

[7] Open Worldwide Application Security Project. (2024). OWASP Top 10 for LLM applications 2025. https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/

[8] Truhn, D., Reis-Filho, J. S., & Kather, J. N. (2023). Large language models should be used as scientific reasoning engines, not knowledge databases. Nature Medicine, 29, 2983–2984. https://doi.org/10.1038/s41591-023-02594-z

[9] Tabassi, E. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

[10] International Organization for Standardization, & International Electrotechnical Commission. (2023a). Information technology- Artificial intelligence-Management system (ISO/IEC Standard No. 42001:2023). International Organization for Standardization. https://www.iso.org/standard/42001

[11] Bellala, K. R. (2025). AI at the edge: Cloud-edge synergy. *International Journal of Innovative Science and Research Technology, 10*(5), 2561–2567. https://doi.org/10.38124/ijisrt/25may967

[12] International Organization for Standardization, & International Electrotechnical Commission. (2023b). Information technology—Artificial intelligence-Guidance on risk management (ISO/IEC Standard No. 23894:2023). International Organization for Standardization. https://www.iso.org/standard/77304.html

[13] Koniakou, V. (2023). From the "why" to the "how" of trustworthy AI: A systematic review on AI governance. AI and Ethics, 3, 629–642.

[14] NIST. (2018). Risk management framework for information systems and organizations: A system life cycle approach for security and privacy (NIST Special Publication 800-37 Rev. 2). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-37r2

[15] Rao, A., Keller, A., Kalra, N., Steed, R., Kwegyir-Aggrey, K., Klyman, K., & Bergman, D. (2026). Challenges to the monitoring of deployed AI systems: Center for AI Standards and Innovation (NIST AI 800-4). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.800-4

[16] Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. In Advances in Neural Information Processing Systems 28 (pp. 2503–2511).

[17] National Institute of Standards and Technology. (2020). NIST privacy framework: A tool for improving privacy through enterprise risk management, Version 1.0. U.S. Department of Commerce. https://doi.org/10.6028/NIST.CSWP.01162020

[18] Bellala, K. R. (2025). Data privacy and compliance in the cloud (GDPR, HIPAA, CCPA). *International Journal of Innovative Science and Research Technology, 10*(7), 2091–2097. https://doi.org/10.38124/ijisrt/25jul1264

[19] Organization for Economic Co-operation and Development. (2019). Recommendation of the Council on Artificial Intelligence (OECD/LEGAL/0449). OECD Legal Instruments. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

[20] 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). Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/

[21] Boeckl, K., & Lefkovitz, N. (2020). NIST privacy framework: A tool for improving privacy through enterprise risk management, Version 1.0 (NIST Cybersecurity White Paper 10). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.CSWP.10

[22] Ross, R. (2018). Risk management framework for information systems and organizations: A system life cycle approach for security and privacy (NIST Special Publication 800-37 Rev. 2). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-37r2

[23] Rao, A., Keller, A., Kalra, N., Steed, R., Kwegyir-Aggrey, K., Klyman, K., Staheli, D., & Bergman, S. (2026). Challenges to the monitoring of deployed AI systems: Center for AI Standards and Innovation (NIST AI 800-4). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.800-4