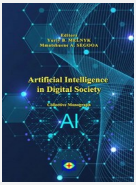
Artificial Intelligence in Digital Society,
Volume 1, 2026
DOI: 10.26697/9786177089192.2026
ISBN 978-617-7089-19-2 (Vol. 1)
ISBN 978-617-7089-18-5 (Series)



Chapter 8. Navigating Governance, Ethics, and Data Security Risks in Artificial Intelligence Adoption

Sathekge M. S. [1], **Bvuma S.** [1]

[1] *University of Johannesburg, South Africa*

Received: *25.10.2025;* ***Accepted:*** *10.02.2026;* ***Published:*** *10.03.2026*

Abstract

The fast and viral uptake of Generative AI (GenAI) and large foundation models (LFMs) in the corporate worlds is a major, but ill-managed, change in organizational security, risk, and governance. Although GenAI involves an overwhelming number of advantages, its implementation creates a new layer of data security threats that traditional ICT security models were not created to cover. The chapter gives a critical review of the situation in governance today and the particular ethical and technical issues that come along with the integration of GenAI. It is analyzed to explain GenAI Security Threats, such as model poisoning and prompt injection, and the highly significant problem of data leakage and exposure of intellectual property (IP). It also explores the Ethical Gaps, which say that inexplicable bias may yield the results of discrimination or generate shadow vulnerability, which could not be audited. The main contribution is the suggestion of a Socio-Technical Governance Framework that incorporates human control, Explainable AI (XAI), and constant security surveillance into the GenAI deployment pipeline. Actionable Best Practices of data sanitization, model validation and defining clear lines of accountability in AI-driven decisions support this framework. This chapter is meant to inform technology leaders and policymakers by expressing the need to have a proactive risk-based approach to ensure GenAI is exploited safely and in a manner that is responsible in the digital society.

Keywords: generative AI, data security, corporate governance, ethical AI, risk management, model poisoning, data leakage explainable AI.

Cite this chapter as:
Sathekge, M. S., & Bvuma, S. (2026). Navigating governance, ethics, and data security risks in artificial intelligence adoption. In Y. B. Melnyk & M. A. Segooa (Eds.), *Artificial Intelligence in Digital Society, Vol. 1*. (pp. 118–131). KRPOCH. https://doi.org/10.26697/aids.2026.8

The electronic version of this chapter is complete. It can be found online in the AIDS Archive https://doi.org/10.26697/aids

Introduction

Context: The Generative AI Discontinuity and Corporate Risk

The adoption of Generative AI (GenAI) and Large Foundation Models (LFMs) into businesses at an accelerated pace and with viral pace can be deemed as one of the largest technological changes of the decade, quickly transforming into an experimental novelty and becoming an essential utility in a business enterprise.

The GenAI is full of potential, and it leads to tangible productivity increases and will change the fundamental operations of code generation, research, and content creation. But because of this pace of adoption driven by business demand of a competitive advantage, it has brought about a major disjuncture in the structure of organizational security, risk, and governance.

The problem is in its structure: GenAI brings a new dimension of data security threats that traditional Information and Communication Technology (ICT) security frameworks were not supposed to deal with (Kendzierskyj et al., 2024). The fact that AI systems are not fully similar to conventional software is based on the fact that threats are no longer targeted at the code but at the very learning and adaptive nature of the model. This requires an interdisciplinary approach that is holistic and in tandem with the future of AI-driven transformation (Radanliev et al., 2025).

Problem Focus and Chapter Goal

A major gap between institutional regulation and the safeguarding of technological innovation and risk has been caused by the rate of GenAI adoption exceeding the maturity of institutional protections. Traditional Information and Communication Technology (ICT) security models were not made to be responsive to the underlying vulnerabilities of probabilistic, content-generating systems (Radanliev et al., 2025).

The chapter of this book argues that this governance gap results in unacceptable corporate exposure in three interrelated areas:

1. Data Security Threats: New attack vectors such as model poisoning and prompt injection compromise model integrity and enable sensitive information to be exfiltrated. This is worsened by the fact that there has been a growing threat of information leakage and intellectual property (IP) exposure as proprietary corporate data are more and more passing to and processing by GenAI tools (Sidorkin, 2025).

2. Ethical and Fairness Gaps: With the complexity and the obscurity of the inner mechanics of the LLMs, it is challenging to analyze the mechanisms to unintentionally reproduce training data-driven biases (Bano et al., 2023; Melnyk, 2025). When applied in the high-stakes corporate security or access control, the outcomes can be discriminatory with shadow vulnerability that cannot be audited in traditional ways.

3. Accountability Deficit: The lack of clear, strictly implemented policies on governance issues surrounding the use of data, along with the self-directed nature

of the work of Generative AI, grossly contributes to the loss of accountability of the locus in cases of error or harm. Such ambiguity significantly increases the chances of the violation of the regulatory requirements (Mandava, 2025).

The Chapter Goal will be the critical synthesis of these challenges and, consequently, to formulate the need to adopt a proactive, risk-based approach to GenAI security governance.

Chapter Contribution and Structure

The fundamental input of this piece of work is the suggestion of a Socio-Technical Governance Framework. This framework is intended to address the governance gap by combining human oversight and the concepts of XAI, including transparency and interpretability, and the ongoing security monitoring as a part of the GenAI deployment pipeline directly. This is a direct way of addressing the causes of the problem of the black box, where the grounding of governance on human-understandable processes is done. The further parts continue with a critical synthesis: initially, the descriptions of the research method and the conceptual basis (Section 2) and the Ethical Gaps Analysis (Section 3) which require systemic change. The chapter then proceeds to introduce the detailed Socio-Technical Governance Framework (Section 4) and the related Actionable Best Practices (Section 5) to be implemented, and finally, is concluded with strategic suggestions on the part of technology leaders and policy makers.

Conceptual Foundation, Research Approach and Methodology

Research Approach and Methodology

The study involves deductive research in which the researcher attempts to derive the topic and the research questions through inductive reasoning. The research employs a deductive research methodology where the investigator tries to come up with the topic and the research questions by using inductive inferences.

The basis of this chapter is a critical synthesis and analysis of authoritative, publicly published reports and a literature review conducted by peer-reviewed sources. The method was chosen to quickly internalise the most recent discoveries, regulatory outlooks, and risk evaluations in regards to GenAI, which tend to be uncovered initially in authoritative industry white papers and specialised, conferences because of the faster rate at which the technology advances.

The search strategy was based on locating the literature published during the past five years (2021-2025), whereby much attention was paid to the publications that were published after 2023 when the GenAI explosion has taken effect after the introduction of LFMs on a large scale. The major search terms were Generative AI, Data Security, Corporate Governance, Ethical AI, Risk Management, Model Poisoning, Data Leakage, XAI. This methodology was important to make sure that the threats found and mitigation measures suggested are extremely relevant and up-to-date.

Conceptual Foundation

According to Chen and Metcalf (2024), Socio-Technical assumes the existence of a company where society and technologies are integrated into a single system. Significant, neither is it possible to conceive the social without the technical, nor the technical without the social. The Socio-Technical Governance Framework, as suggested is deeply based on the Socio-Technical Systems (STS) theory. STS aims to regard organizations, processes, and technology as not separate entities, but as one system. This is necessary within the framework of the Digital Society since the governance failures of the GenAI are not a matter of technical deficiency (e.g., the deficiency of the code) but the deficiency of the systemic engagement between the algorithmic functionality and the human policy, oversight, and culture.

Using an STS lens offers the conceptual rationale behind the structure of the framework that requires convergence of:

- *Technical Pillar*: The introduction of techniques such as XAI that deal with the technical black box issue by introducing transparency and interpretability.

- *Social Pillar*: Implementing the human control and obligatory forms of accountability including the governance boards, auditor trails that would verify the compliance of AI-made decisions to the ethical and regulatory policy.

The framework therefore maintains that addressing the risks of model poisoning and bias cannot be achieved only by methods of improved filtering algorithms, and it is important to bring clarity in human roles and enforceable policies in terms of access to data and decision review which entrench the principle that technology and human action are mutually constitutive in the determination of a security and ethical outcome. To account for the results of any technology, including AI it is necessary to concentrate on the in-between space between these two pillars, which is also complicated (Chen and Metcalf, 2024).

GenAI Security Threats and Technical Risks

Threats to Model Integrity

GenAI systems are compromised with advanced attacks that attack the data used to train the model and query information.

A. Model Poisoning

Model poisoning refers to malice inputs or maliciously labeled data to the training pipeline to cause the model to adopt tainted behavior. This is a direct contravention of integrity and reliability of the model.

- Targeted Attacks (Backdoors): A backdoor attack is a branch of data poisoning in which the malicious activity is not triggered unless a certain trigger is met or a specific phrase is used (Souly, et al., 2025). These attacks are meant to cause the model to misbehave under a certain trigger that is not explicit in the input, but overall functions well. This is what makes the attack difficult to notice when going through regular validation. According to research, a minimum of 250 documents can be used to reliably backdoor LLMs, despite the overall size of the model itself (Souly, et al., 2025).

- Non-Targeted Attacks (Availability): In such attacks, big amounts of noise data or incorrect data are injected into the overall model and affect its functionality and resilience. It is aimed at leading to massive performance decrease, which causes inaccurate outputs as well as a higher error rate, which jeopardizes the utility and credibility of the model (Hubinger, et al., 2024).

B. Prompt Injection

This is an immediate injection where malicious input is developed by attackers to either disrupt or disorient an AI system. Prompt Injection is the highest-ranked security risk, which is classified as one of the most severe vulnerabilities in the applications of LLM. It is the type of attack in which the developers of the original system provide maliciously designed, harmful inputs to the model, and they override the original system instructions or system prompt. (Redbot Security, 2025).

- Mechanism: Prompt injection takes advantage of the fact that the LLM has no reliable way of differentiating trusted system code and untrusted data, which are sent in by a user. (Sidorkin, 2025).

- Consequences: An effective attack may force the model to do a malicious act e.g. coughing confidential information, cross-site scripting, or unwanted code (Bowen et al., 2025). This is especially hazardous in systems that rely on Retrieval-Augmented Generation (RAG), in which the prompts can trigger the model to retrieve and disclose sensitive information in internal, and also a vector-database knowledge sources (Redbot Security, 2025; Souly, et al., 2025).

Threats to Data and Confidentiality

GenAI as a concept poses extreme risks to corporate information confidentiality and intellectual property (IP), and gives rise to the liability of non-regulatory compliance (Ranjan and Kettani, 2025).

A. Data Leakage and Intellectual Property (IP) Exposure

One of the most pressing concerns is the accidental utilization of the proprietary information to train the models that are public or unsecured (Sidorkin, 2025):

- Insecure Data Ingress/Egress: There has not been a defined, enforced governing policy concerning how proprietary data is inputted into GenAI tools (ingress) or how outputs are managed (egress) that is intolerable.

- Model Memorization: Models can unintentionally memorize certain data in the training corpus during training. In the case of this memorized data, this may be hacked by attackers using certain query methods in case it is proprietary or a source of Personally Identifiable Information (PII).

- Shadow AI Risk: Unmanaged employees use public GenAI services the exposure to unknown and uncharted IP risk on the organization directly through employees who feed the model with confidential data to perform tasks such as summarization or code completion.

B. Regulatory Consequences

These technical failures are converted into direct regulatory risks. The absence of strong security protocols with regard to the management of data is a direct conflict with the new global AI standards and data protection regulations. C leakage of the sensitive information on the grounds of model compromise may result in both expensive regulatory non-conformity and grave reputational losses (Chesterman, 2025).

Ethical Gaps: Bias, Fairness, and Accountability

The technical risks of GenAI are associated with deep ethical loopholes of bias, fairness and accountability. These present regulatory risks, which normally harm reputation and trust more than breaches.

The Challenge of Unexplainable Bias

The major ethical issue of the deployed GenAI is the expression of the bias based on the training data. Such systemic weaknesses can be promoted and exaggerated by models that are trained on biased, historically prejudiced, or unrepresentative data and in their outputs and decisions (Radanliev et al., 2025; Ranjan and Kettani, 2025).

- Bias in Security Decision-Making: When GenAI becomes part of a high-stake system, say, threat detection, access control, or employee monitoring, any bias, even when not explainable, may cause discriminatory results, including the marginalization of certain demographics.
- The Opacity Problem: LFMs are often complex systems whose decision-making processes are so opaque that it makes them black boxes. It is so invisible that the human operators or auditors cannot effectively mitigate it since it is incredibly hard to establish the reason behind a specific decision or how a bias is introduced. Such transparency will compromise the fundamental values of fairness and equity in operations (Mandava, 2025).

The Accountability Deficit and Shadow Vulnerabilities

This does not mean that the responsibility is absent, a problem with GenAI is that its inherent opaqueness always creates a lack of accountability, making it difficult to distinguish who should take responsibility and who should not in a situation where a system delivers a biased or damaging result.

Accountability in a business setting entails being traceable and assigning responsibility to algorithmic activities (Mersah et al., 2025; Ranjan & Kettani, 2025; Sidorkin, 2025).

These can lead to:

- Diffused Responsibility: It is hard to figure out who should take responsibility over autonomous GenAI decisions: the author of the decision is the developer, the curator, the team, or the supervisor? (Janssen, 2025).
- Shadow Vulnerabilities: The lack of accountability give rise to shadow vulnerabilities that are invisible risks that are systemic and cannot be audited. They pose the threat of continuous damages and non-compliance with regulations. The governance should entail accountability at the board level of all GenAI risks (Mandava, 2025; Sidorkin, 2025).

The Imperative for Explainable AI (XAI)

The need to integrate XAI is essential so that responsible governance is achieved, and technical tools are offered to address the issues of connecting the incomprehensible model results to human understanding (Mandava, 2025).

These are:

- Achieving Transparency and Interpretability: Creating Transparency and Interpretability XAI algorithms such as SHAP (quantifies feature contribution) or LIME (local surrogate models) can offer human-understandable components to individual AI decisions (Shankar, 2025). This will enable security analysts to verify alerts as well as ethics officers debugging models before and after deployment.
- Supporting Accountability: XAI contributes to antecedence and auditability of results through clear-cut evidence and supports the Socio-Technical Governance Framework in aspects of integrating technical transparency and human controls (Chen and Metcalf, 2024).

The Socio-Technical Governance Framework

The technical and ethical loopholes require an alternative approach to governance other than conventional policy. To cope with the complexity of GenAI, this chapter suggests a Socio-Technical Governance Framework that would combine human regulation with system transparency to handle systemic failures.

The Rationale

The main objective of this framework is to create a binding connection between corporate policy (the social pillar) and algorithmic functionality (the technical pillar).

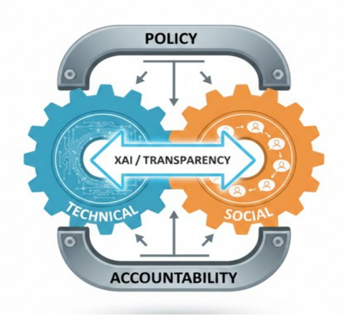
The framework is designed in such a way that continuous monitoring, explainability and human judgment are incorporated into the lifecycle of GenAI deployment, not disengaged compliance inspections. The given practice is essential to developing reliable AI in stakes-based settings (Chen and Metcalf, 2024).

Framework Presentation

Figure 8.1 represents the operational scheme of the Socio-Technical Governance Framework.

Figure 8.1

Socio-Technical Governance Framework



Tables 8.1-8.4 illustrate the specific elements of the framework core, linkages and pillars.

Central Core: The GenAI Lifecycle

This is the point where governance must be applied.

Table 8.1

The Gen AI Lifecycle

Component	Description	Relevance
GenAI Deployment Pipeline (Policy & Accountability)	The central process encompassing model training, deployment, inference, and operation	Represents the system being governed; all risks (poisoning, injection) occur here

The Linkages (The Mechanisms)

These components span both pillars and are necessary for communication and transparency.

Table 8.2

The Mechanism

Component	Description	Technical Pillar Function	Social Pillar Function
Explainable AI (XAI) Principles	Implementation of interpretability tools (e.g., LIME, SHAP).	Provides technical transparency and helps debug bias.	Enables human validation and justification of AI-driven decisions.
Transparency and Continuous Security Monitoring	Real-time monitoring of model inputs, outputs, and performance.	Detects and flags adversarial attacks (prompt injection, data leakage).	Triggers human oversight, remediation, and reporting to the Governance Board .

The Technical Pillar (Mitigation and Transparency)

This pillar focuses on the system's defensive and diagnostic capabilities.

Table 8.3

The Mitigation and Transparency

Component	Description	Goal / Mitigation	Source
Data Sanitization and Vetting	Strict policies on data ingress; use of anonymization and source checks.	Mitigates Model Poisoning and Data Leakage.	Mandava (2025)
Layered Input/Output Filtering	Mechanisms to sanitize user prompts (input) and vet model responses (output).	Mitigates Prompt Injection and prevents unauthorized data exfiltration.	Sidorkin (2025)
Adversarial Robustness Testing	Mandatory Red Teaming and stress testing.	Proactive identification of vulnerabilities before deployment.	Bowen, et al. (2025)
Immutable Audit Trails	Technical mechanism to log all AI decisions and associated XAI rationale.	Supports the Accountability principle.	Mersah, et al. (2025)

The Social Pillar (Oversight and Accountability)

This pillar focuses on the institutional and human structures that enforce ethical policy and ensure compliance.

Table 8.4

The Oversight and Accountability

Component	Description	Goal / Mitigation	Source
AI Governance Board (AGB)	Cross-functional executive committee (Legal, Ethics, IT, Senior Management).	Sets high-level policy, conducts risk assessments, and retains ultimate decision authority.	Taeihagh (2025)
Defined Human Oversight Roles	Clear assignment of responsibility (e.g., Human-in-the-Loop, Human-over-the-Loop).	Addresses the Accountability Deficit and prevents discriminatory outcomes.	Kandikatla & Radeljić (2025)
Ethics and Remediation Policy	Formal protocol for pausing, remediating, or retiring a GenAI system.	Ensures systems align with corporate fairness and ethical commitments.	Chesterman (2025)
Regulatory Compliance & Reporting	Formal processes for meeting standards (e.g., POPIA, GDPR, EU AI Act).	Mitigates Regulatory Non-Compliance risk.	Radanliev et al. (2025)

Actionable Best Practices for Implementation

This part delivers the practical guide that was promised in the abstract giving steps on how to implement it. The Socio-Technical Governance Framework requires the shift to the realms of theory and best practices that are warehoused on a real-life agenda, and these practices should encompass the implementation of security and ethics into the Generative AI (GenAI) lifecycle. Such practices are the working units of the framework providing technical strength and institutional responsibility against the mentioned dangers of poisoning, injecting, and losing the data.

Securing the Data Pipeline (Mitigating Poisoning/Leakage)

The initial defence against Model Poisoning and Data Leakage/IP Exposure is protecting the data by which the training process and inference is done.

The practices are aimed at protecting the ingress (input) and egress (output) points of the data:

- Data Sanitization and Minimization: Organizations need to strictly screen and sanitise all training data considering the whole corpus as potentially untrusted. This includes implementing data minimization, processing only the data required to train, and applying such methods as tokenization and data masking to replace or cover Personally Identifiable Information (PII) and sensitive Intellectual Property (IP) (OWASP, 2024; Sidorkin, 2025).

- Source Vetting and Continuous Integrity Checks: To mitigate the possibility of Model Poisoning, third-party data sources as well as internal data sources should be regularly checked by integrity tests and formally vetted. It is a method that protects against the introduction of malicious samples, which can be performed with high precision and hides silently, reliably breaking models of any scale (Souly, et al., 2025; Sidorkin, 2025).

- Enforced Data Ingress/Egress Policies: There should be explicit policies of governance that determine how proprietary data can be ingested into GenAI tools (ingress) and how sensitive summary or code can be egressed (egress). This directly prevents the threat of Shadow AI and unintentional IP leakage. (OWASP, 2024; Mandava, 2025).

Model Validation and Adversarial Testing (Mitigating Prompt Injection)

To make the technical pillar of the framework solid, the persistent validation and testing cannot be limited to the common functional quality assurance but should consider the special weak points of GenAI.

- Layered Input Validation (Prompt Injection Defense): Any input by the user should be considered untrusted. This should be implemented in a layered manner with both rule based filters and AI based classifiers that can both identify and neutralise malicious or obfuscated instructions before they can reach the core LLM. It is the most significant technical defense against timely injection. (OWASP, 2024; Taeihagh, 2025).

- Continuous Adversarial Robustness Testing (Red Teaming): Organizations will have to institutionalize focused Red Teaming actions, such that security specialists simulate adversarial attacks, in particular, the ability of the model to resist sophisticated prompt injection and data exfiltration attacks. This constant validation procedure keeps the defenses up to date with the changing attack vectors (Radanliev et al., 2025).
- Output Filtering and Sanitization: The response generated by the model should be inspected to detect some evidence of illicit information e.g. confidential internal information, PII or malicious instructions. This is a final safeguard against data leakage resulting from a successful prompt injection attack (Sidorkin, 2025).

Operationalizing Accountability and Oversight

These practices codify the social pillar of the framework by creating the clear lines of responsibility required to handle the ethical risk and compliance with regulations.

- Establish Mandatory Audit Trails: One of the conditions that is not negotiable is to create immutable and auditable records of every decision, classification, or action made depending on the GenAI system. This technical necessity is a prerequisite to empower the humanistic value of traceability and accountability, where all the results are justifiable and checked (Radanliev et al., 2025).
- Define and Enforce Lines of Responsibility: The AI Governance Board should establish and implement obvious lines of responsibility, clarifying the ultimate accountable human manager or executive to the consequences of a deployed GenAI system. This minimizes the risk of diffused responsibility in the case where systems are unintentionally biased in any way or that they yield discriminatory results (Kandikatla and Radeljić, 2025; Janssen, 2025).
- Mandate Remediation Mechanisms: The organization should specify transparent, tested standards to shut down, remediate, or discontinue an AI system the moment its ongoing security monitoring or XAI analysis of the result reveals that it is introducing or strengthening bias, modifying data integrity or breaching ethical policy. This offers a required safety valve to ensure the public trust and avoid disastrous breakdowns (Ranjan & Kettani, 2025; Taeihagh, 2025).

Conclusion and Recommendations

As can be seen by the results of this chapter, the risks presented by Generative AI (GenAI) are not just technical issues but governance failures in their entirety. In order to ensure that the innovative power of GenAI will be used safely and responsibly, technology leaders and policymakers should take a risk-based proactive approach. The proposed Socio-Technical Governance Framework contains the following recommendations, which are based on the regulatory governance of Digital Society.

Recommendations for Technology Leaders (Implementers)

1. Mandate Explainable AI XAI Integration for High-Risk Systems: Leadership should mandate XAI integration of high-stakes GenAI security deployments to offer transparency, which can allow human verification and debugging of bias to overcome accountability shortcomings.
2. Institutionalize Adversarial Testing: Red Teaming and adversarial stress testing should be institutionalized and continue continually. This is a necessary practice to detect dynamic vulnerabilities to timely inject and maintain model resilience to model poisoning.
3. Enforce Immutable Audit Trails: To realize accountability, technology departments should deploy technologies that generate immutable auditable records of every decision made or impacted by GenAI. This will make any ethical or security lapses attributable to certain activities in support of the social pillar of accountability.
4. Implement Zero-Trust Data Ingress/Egress Policies: Technology leaders need to have and enforce express policies on data ingress and egress using GenAI tools, such as using data sanitization techniques and Data Loss Prevention (DLP) to reduce the possibility of data leakage and IP exposure.

Recommendations for Policymakers (Regulation and Oversight)

1. Establish Clear Board-Level Accountability: The policies should specify accountability of the results of high-risk AI systems to a particular executive or a board-level committee. This eliminates diffusion of responsibility and also makes sure that risks of not adhering to the regulations are met at the topmost level.
2. Harmonize XAI Requirements: To achieve algorithmic fairness and explainability, policymakers ought to employ minimal requirements on XAI interpretability in the regulated industries. These standards need specifically to deal with the way that companies need to show that their systems of GenAI have been tested and fixed in case of unexplainable bias.
3. Mandate Transparency in Data Provenance Regulatory frameworks ought to demand more disclosure around the provenance and composition of the training data applied by LFMs in the effort to assist organizations to alleviate the threat of model poisoning and shadow vulnerabilities.

Conclusion

This chapter has critically examined the GenAI governance landscape, and it has been shown that the adoption of LFMs has occurred at a rapid, heavily viral pace, which resulted in a considerable discontinuity in the organization security, risk and ethics. The review established that the current ICT security models are poorly prepared to address the particular threats of model poisoning and timely injecting, the lack of transparency in these systems generates an ethical dilemma of bias, fairness and accountability. The main contribution of the given work is the suggested Socio-Technical Governance Framework. This model offers a solid framework of controlling the unavoidable merging of human resourcefulness, business data, and machine generated innovation. The framework guarantees that the ethical standards and security protocols are directly embedded in the deployment life-cycle, through a systematic approach to the combination of the Technical Pillar (XAI and continuous monitoring) and the Social Pillar (human oversight and defined accountability). Such an active and risk-sensitive stance is required to reduce the risk of the high cost of non-compliance with regulatory policies and reputation losses, which means that GenAI will be incorporated safely and responsibly into the new digital society.

References

Bano, M., Zowghi, D., Shea, P., & Ibarra, G. (2023). *Investigating responsible AI for scientific research: An empirical study*. arXiv. https://doi.org/10.48550/arXiv.2312.09561

Bowen, D., Murphy, B., Cai, W., Khachaturov, D., Gleave, A., & Pelrine, K. (2025). Scaling trends for data poisoning in LLMs. *Proceedings of the AAAI Conference on Artificial Intelligence*, *39*(26), 27206–27214. https://doi.org/10.1609/aaai.v39i26.34929

Chen, B. J., & Metcalf, J. (2024, May 28). *Explainer: A sociotechnical approach to AI policy*. Data & Society Research Institute. https://datasociety.net/library/a-sociotechnical-approach-to-ai-policy/

Chesterman, S. (2025). Good models borrow, great models steal: Intellectual property rights and generative AI. *Policy and Society*, *44*(1), 23–37. https://doi.org/10.1093/polsoc/puae040

Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., ... & Maxwell, T. C. (2024). *Sleeper agents: Training deceptive LLMs that persist through safety training*. arXiv. https://doi.org/10.48550/arXiv.2401.05566

Janssen, M. (2025). Responsible governance of generative AI: Conceptualizing GenAI as complex adaptive systems. *Policy and Society*, *44*(1), 38–51. https://doi.org/10.1093/polsoc/puae039

Kandikatla, L., & Radeljić, B. (2025, October 10). *AI and human oversight: A risk-based framework for alignment*. arXiv. https://doi.org/10.48550/arXiv.2510.09090

Kendzierskyj, S., Jahankhani, H., & Hussien, O. (2024). Space governance frameworks and the role of AI and quantum computing. In H. Jahankhani (Ed.), *Space law and policy* (pp. 1–39). Springer. https://doi.org/10.1007/978-3-031-62228-1_1

Mandava, S. (2025). Explainable data governance using XAI techniques to enhance traceability, transparency, and accountability in AI systems. *Applied Data Science and Analysis*, *2025*(1). https://doi.org/10.58496/ADSA/2025/001

Melnyk, Y. B. (2025). Should we expect ethics from artificial intelligence: The case of ChatGPT text generation. *International Journal of Science Annals, 8*(1), 5–11. https://doi.org/10.26697/ijsa.2025.1.5

Mersah, M. A., Yigezu, M. G., Tonja, A. L., Shakil, H., Iskander, S., Kolesnikovs, O., & Kalita, J. (2025). Explainable AI: XAI-guided context-aware data augmentation. *Expert Systems with Applications*, *289*(128364). https://doi.org/10.1016/j.eswa.2025.128364

OWASP. (2024). *OWASP top 10 for LLM applications 2025*. https://owasp.org/www-project-top-10-for-large-language-model-applications/assets/PDF/OWASP-Top-10-for-LLMs-v2025.pdf

Radanliev, P., Santos, O., & Ani, U. D. (2025). Generative AI cybersecurity and resilience. *Frontiers in Artificial Intelligence*, *8*(1568360), 1–18. https://doi.org/10.3389/frai.2025.1568360

Ranjan, R. P., & Kettani, Z. (2025). Scenario planning for managing AI disruption risk: A 3C-AI framework. *California Management Review*. https://cmr.berkeley.edu/2025/10/scenario-planning-for-managing-ai-disruption-risk-a-3c-ai-framework/

Redbot Security. (2025, October 30). *Prompt-injection-attacks-ai-security-2025*. https://redbotsecurity.com/prompt-injection-attacks-ai-security-2025/

Shankar, V. (2025). Machine learning for Linux kernel optimization: Current trends and future directions. *International Journal of Computer Sciences and Engineering*, *13*(3), 56–64. https://doi.org/10.26438/ijcse/v13i3.5664

Sidorkin, A. M. (2025). AI platforms security. *AI-EDU Arxiv*, *2025*(1). https://journals.calstate.edu/ai-edu/article/view/5444

Souly, A., Rando, J., Chapman, E., Davies, X., Hasircioglu, B., Shereen, E., ... & Kirk, R. (2025, October 8). *Poisoning attacks on LLMs require a near-constant number of poison samples*. arXiv. https://doi.org/10.48550/arXiv.2510.07192

Taeihagh, A. (2025). Governance of generative AI. *Policy and Society*, *44*(1), 1–22. https://doi.org/10.1093/polsoc/puaf001

Tedeneke, A. (2023, June 26). *World Economic Forum launches AI Governance Alliance focused on responsible generative AI*. World Economic Forum. https://www.weforum.org/press/2023/06/world-economic-forum-launches-ai-governance-alliance-focused-on-responsible-generative-ai/

Information about the authors:

Sathekge Machiniba Sylvia – https://orcid.org/0009-0001-9410-3267; Doctor of Business Administration, Doctor, Professor of Practice, University of Johannesburg, Johannesburg, South Africa.

Bvuma Stella – https://orcid.org/0000-0001-8351-5269; PhD in Information Technology Management; Professor, Director, University of Johannesburg, Johannesburg, South Africa.