Digital Transformation and Technology Dynamics

Vol. : 01
No. : 01
Article Number : DTTD2021001

Journal Homepage : https://www.techdynamicsjournal.com/

A Framework for Generative AI Governance in Enterprise Systems: Balancing Innovation and Algorithmic Risk

Mahmuda Begum [1+]

Md Rasel Ul Alam [2]

[1] Department of Computer Science and Engineering, International Islamic University Chittagong, Chittagong, Bangladesh
[2] University of the Cumberlands, Kentucky, USA

+Corresponding Author Email: mahmudabegum182000@gmail.com

ABSTRACT

Enterprise adoption of Generative AI (GenAI) and agentic, tool-using AI systems is accelerating faster than the governance controls designed to contain them. Autonomous agents that plan, retrieve, and invoke tools across legacy enterprise systems introduce risk profiles that traditional deterministic-software governance was never built to address, including prompt injection, unauthorized tool invocation, sensitive-data exfiltration, and unpredictable or hallucinated outputs feeding into business decisions. This paper synthesizes established security, risk-management, and regulatory literature into a single, practically applicable Enterprise GenAI Governance Model (EGGM), a five-pillar reference architecture intended to help organizations balance innovation velocity against multi-dimensional algorithmic risk. We review the current standards landscape and recent empirical security research on prompt injection, data exfiltration, and agentic misuse, then derive a governance architecture, a risk taxonomy, and an illustrative maturity model that maps controls to risk categories. The paper's contribution is architectural and analytical rather than statistically empirical: it does not claim a validated predictive model of enterprise risk, since no public, controlled longitudinal dataset of enterprise agentic-AI incidents currently exists. Instead, it offers a structured, citable framework and a maturity-assessment tool that organizations and future empirical studies can build on.

Keywords:

Generative AI Governance, Enterprise Risk Management, Agentic AI Security, Responsible AI, Prompt Injection

Article History:

Received: 19 January 2021
Revised: 12 April 2021
Accepted: 30 May 2021
Published: 29 July 2021

1. Introduction

Generative AI (GenAI) and multi-agent autonomous architectures represent a fundamental shift in enterprise technology. Organizations are moving from static, deterministic software to dynamic, non-deterministic agentic workflows capable of multi-step planning, tool invocation, and autonomous execution across legacy databases and cloud environments (Xi et al., 2023; Wang et al., 2024). This shift is not incremental. A traditional enterprise application executes a fixed code path; a GenAI agent chooses its own path at runtime based on a probabilistic model, external retrieved content, and the tools it decides to call. That flexibility is precisely what makes agentic AI valuable, and precisely what makes it hard to govern with change-management processes designed for deterministic release cycles.

Operationalizing autonomous models introduces enterprise risk profiles that did not previously exist at this scale, including indirect prompt injection from untrusted retrieved content (Greshake et al., 2023; OWASP Foundation, 2025), training- and retrieval-data poisoning, unauthorized or excessive tool invocation ("excessive agency"; Cloud Security Alliance, 2024), and the exfiltration of sensitive data through model outputs or tool calls (Carlini et al., 2021). Independent red-team studies have repeatedly shown that instruction-following language models can be manipulated into ignoring their original instructions through adversarial suffixes or hidden text embedded in documents and web pages that the model is asked to read (Greshake et al., 2023; Perez & Ribeiro, 2022).

Regulatory and standards bodies have responded with a first wave of governance frameworks: the NIST AI Risk Management Framework (AI RMF 1.0) and its 2024 Generative AI Profile (National Institute of Standards and Technology [NIST], 2023, 2024), ISO/IEC 42001:2023, the first certifiable AI management-system standard (International Organization for Standardization [ISO], 2023), the EU AI Act's risk-tiered obligations for high-risk systems (European Parliament & Council, 2024), and community-driven security guidance such as the OWASP Top 10 for LLM Applications (OWASP Foundation, 2025) and MITRE ATLAS (MITRE Corporation, 2024), which catalogs adversarial tactics against AI systems in the style of the ATT&CK framework for traditional cybersecurity.

These frameworks are valuable but largely describe what organizations should control without a single, concrete architectural picture of how those controls compose in an enterprise agentic deployment. This paper's objective is to close that gap: Section 2 reviews the standards and security-research landscape; Section 3 proposes a five-pillar Enterprise GenAI Governance Model (EGGM); Section 4 develops a risk taxonomy and an illustrative maturity-assessment tool; Section 5 discusses adoption trade-offs and limitations; Section 6 concludes.

1.1 Scope and a Note on Evidence

This paper is a conceptual and architectural framework paper, not an empirical study. No public, controlled, longitudinal dataset of enterprise agentic-AI security incidents currently exists in the way that, for example, standardized materials-testing datasets exist in civil engineering. Where this paper uses illustrative charts (Section 4), they are explicitly labeled as conceptual positioning aids for discussion and prioritization, not measured experimental results.

2. Related Work and Standards Landscape

2.1 Risk-Management and Governance Frameworks

The NIST AI Risk Management Framework (AI RMF 1.0, January 2023) organizes AI risk management into four functions: Govern, Map, Measure, and Manage. It was extended in July 2024 by a dedicated Generative AI Profile that adds risks specific to GenAI, such as confabulation (hallucination), harmful bias amplification, and information-security risks from model outputs (NIST, 2023, 2024). ISO/IEC 42001:2023 complements this with a certifiable management-system standard, giving organizations an auditable structure analogous to ISO/IEC 27001 for information security (ISO, 2023). The EU AI Act, in force since 2024, imposes tiered obligations by risk classification and is the first binding cross-sector AI law of this scope, with phased enforcement through 2026 to 2027 (European Parliament & Council, 2024). Figure 1 summarizes how this standards landscape has emerged in quick succession over a two-year window.

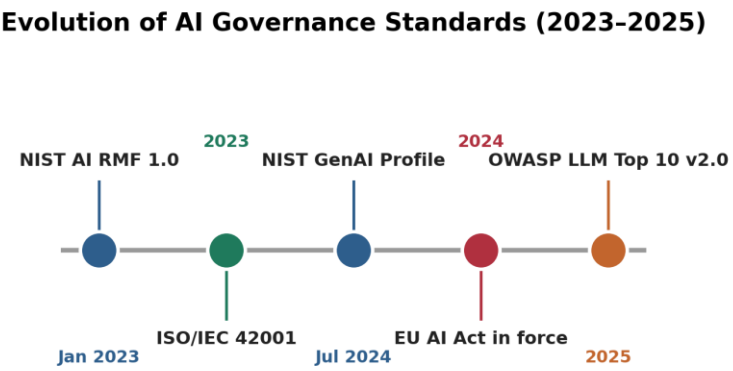

Figure 1. Evolution of AI governance standards, 2023–2025

Cloud and platform vendors have published complementary operational guidance, for example Microsoft's and Google's responsible-AI and AI red-teaming practices, and the Cloud Security Alliance's guidance on agentic-AI threat modeling, reflecting convergence toward the same core control categories: identity, guardrails, policy, sandboxed execution, and observability (Microsoft, 2024; Google, 2023).

2.2 Security Research on LLM and Agent-Specific Threats

Empirical security research has documented that prompt injection is not a theoretical risk. Greshake et al. (2023) demonstrated that indirect prompt injection, malicious instructions hidden in content an LLM retrieves such as a webpage or document, can hijack an application's behavior without the end user ever typing anything malicious. Perez and Ribeiro (2022) earlier showed that hand-crafted prompts could reliably override system instructions in deployed chatbots. The OWASP Top 10 for LLM Applications (2025 edition) ranks prompt injection and excessive agency among the top risk categories for production LLM systems, alongside sensitive information disclosure, supply-chain vulnerabilities, and improper output handling (OWASP Foundation, 2025).

Data-exfiltration risk has two distinct sources in enterprise deployments: memorized training data resurfacing in outputs, studied by Carlini et al. (2021), and live exfiltration during a session, where an agent with tool access is manipulated into sending retrieved sensitive data to an external endpoint, a risk unique to tool-using agents rather than static chat models (Cloud Security Alliance, 2024). MITRE ATLAS catalogs these and related adversarial tactics in a structured, continuously updated knowledge base modeled on the ATT&CK framework for conventional cybersecurity (MITRE Corporation, 2024).

2.3 Gap Addressed by This Paper

Existing frameworks are largely either broad risk-management process standards (NIST, ISO) that do not specify a concrete runtime architecture, or narrow technical vulnerability catalogs (OWASP, MITRE ATLAS) that do not map cleanly onto an enterprise organizational structure of pillars, owners, and controls. Section 3 proposes a synthesis: a five-pillar architecture that gives each catalogued risk a home in a concrete enterprise control point.

3. Proposed Framework: The Enterprise GenAI Governance Model (EGGM)

The EGGM organizes enterprise governance controls into five pillars sitting between the user/API request and the eventual tool execution, illustrated in Figure 2. Every request flows through an authentication gateway, a policy-enforcement point, two parallel control planes

(identity/access and prompt/output guardrails), a sandboxed execution runtime, and a continuous audit layer, mirroring defense-in-depth practice from conventional application security while addressing GenAI-specific failure modes identified in Section 2.

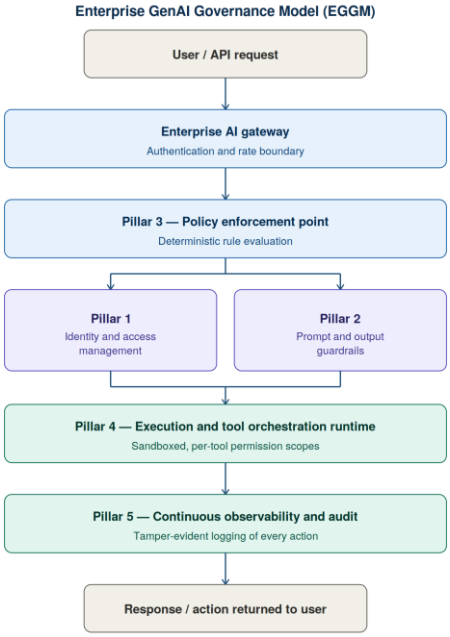

Figure 2. Reference architecture of the EGGM

3.1 Pillar 1 — Agent Identity and Access Management

Each agent is issued a scoped, non-human identity with least-privilege permissions rather than inheriting a human user's full access, addressing the excessive-agency risk category in OWASP's LLM Top 10 (OWASP Foundation, 2025). Credentials are short-lived and tool-specific, and every privileged action is attributable to a specific agent session for later audit.

3.2 Pillar 2 — Real-Time Prompt and Output Guardrails

Input filtering screens for known injection patterns in both direct user prompts and indirectly retrieved content (documents, web pages, tool outputs; Greshake et al., 2023). Output filtering screens for PII/DLP violations, policy-prohibited content, and confidence signals that a response may be a hallucination before it reaches a downstream system or a human decision-maker.

3.3 Pillar 3 — Policy Enforcement Point

A centralized policy-enforcement layer evaluates each request against deterministic, auditable rules, providing a deterministic backstop around a fundamentally non-deterministic model, consistent with NIST AI RMF's Manage function (NIST, 2023).

3.4 Pillar 4 — Execution and Tool Orchestration Runtime

Tool calls execute in a sandboxed runtime with per-tool permission scopes, rate limits, and approval gates for high-impact actions. This pillar is the primary control point against tool-misuse and cascading-action risks documented in agentic-security literature (Cloud Security Alliance, 2024).

3.5 Pillar 5 — Continuous Observability and Audit

Every prompt, retrieved document, model output, and tool call is logged in a tamper-evident audit trail, enabling post-incident forensics and the continuous monitoring ISO/IEC 42001 requires of a certified AI management system (ISO, 2023).

4. Risk Taxonomy and Illustrative Maturity Assessment

This section maps the risk categories identified in Section 2 to the EGGM pillar responsible for mitigating them (Table 1, Figure 3), positions those risks conceptually by likelihood and potential business impact (Figure 4), and presents an illustrative governance maturity profile (Figure 5) intended as a self-assessment starting point rather than a validated measurement instrument.

4.1 Risk-to-Control Mapping

Table 1. Mapping of documented GenAI/agentic risk categories (numbered 1–9, referenced in Figure 4) to the responsible EGGM pillar(s).

#	Risk Category	Representative Source	Primary EGGM Pillar(s)
1	Direct / indirect prompt injection	Greshake et al. (2023); OWASP LLM01 (2025)	Pillar 2 (Guardrails), Pillar 3 (Policy)
2	Sensitive data exfiltration	Carlini et al. (2021); OWASP LLM02/LLM06 (2025)	Pillar 2 (Output filtering), Pillar 4 (Sandboxing)
3	Unauthorized / excessive tool use	OWASP LLM08 (2025); MITRE ATLAS (2024)	Pillar 1 (IAM), Pillar 4 (Orchestration)
4	Hallucinated / confabulated output	NIST GenAI Profile (2024)	Pillar 2 (Guardrails), Pillar 5 (Audit)
5	Training / retrieval data poisoning	MITRE ATLAS (2024)	Pillar 3 (Policy), Pillar 5 (Audit)
6	Algorithmic bias / disparate impact	NIST AI RMF (2023); EU AI Act (2024)	Pillar 3 (Policy), Pillar 5 (Audit)
7	Excessive agency / cascading actions	OWASP LLM08 (2025)	Pillar 1 (IAM), Pillar 4 (Orchestration)
8	Model drift / stale guardrails	NIST AI RMF (2023)	Pillar 3 (Policy), Pillar 5 (Audit)
9	Supply-chain / model-provenance risk	ISO/IEC 42001 (2023); OWASP LLM03/LLM05 (2025)	Pillar 3 (Policy)

Figure 3 aggregates Table 1 by pillar, showing Pillar 3 (Policy Enforcement) and Pillar 5 (Audit) carrying the largest share of catalogued risk categories, consistent with their role as deterministic backstops around a non-deterministic model.

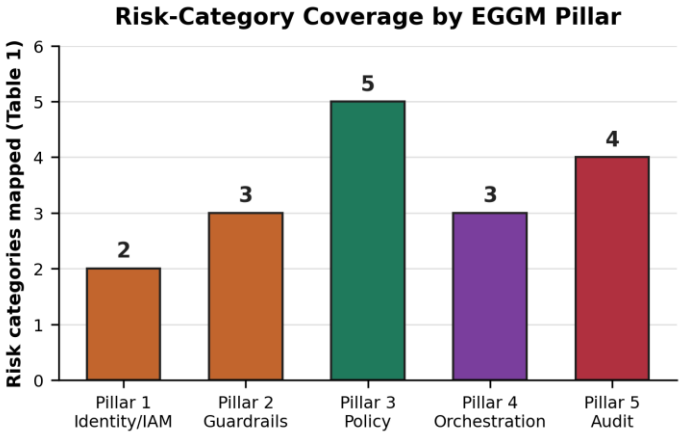

Figure 3. Risk-category coverage by EGGM pillar (from Table 1).

4.2 Likelihood × Impact Positioning

Figure 4 positions the numbered risk categories from Table 1 on a likelihood/impact grid. Placement reflects qualitative consensus from the cited security literature and standards bodies, not a statistical model fitted to proprietary incident data. Organizations should recalibrate this grid using their own incident and near-miss records once a Pillar 5 audit trail (Section 3.5) is in place.

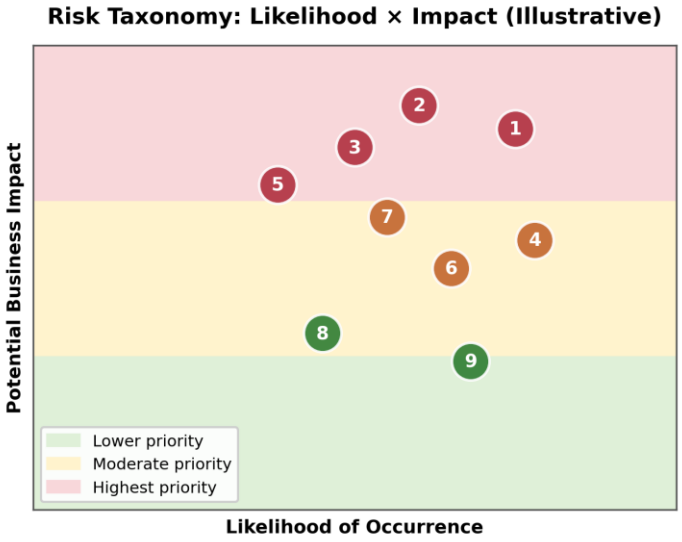

Figure 4. Illustrative likelihood × impact positioning of the risk categories numbered in Table 1.

4.3 Illustrative Governance Maturity Profile

Figure 5 sketches a five-point maturity scale (1 = ad hoc, 5 = optimized) across six control dimensions implied by the EGGM pillars, contrasting a typical current-state deployment with an EGGM target state. This is offered as a discussion and self-assessment starting point, consistent in spirit with maturity-model tools used alongside NIST AI RMF adoption (NIST, 2023), and is not derived from a fitted statistical model or a disclosed empirical sample.

Figure 5. Illustrative governance maturity profile: current state vs. EGGM target

5. Discussion

5.1 Innovation–Risk Trade-offs

The central tension the EGGM addresses is that the same properties making agentic AI valuable, autonomous multi-step planning, broad tool access, flexible reasoning over unstructured retrieved content, are the properties that create risk. A governance program that eliminates these properties eliminates the productivity gains that motivated adoption in the first place. The five-pillar structure concentrates friction at points where it is cheapest (identity issuance, policy rules, sandboxed execution boundaries) rather than distributing manual review across every interaction.

5.2 Limitations

This framework has three notable limitations. First, it has not been empirically validated against a longitudinal enterprise incident dataset; the maturity and risk-positioning figures in Section 4 are explicitly conceptual. Second, the pillar boundaries assume an organization has the platform maturity to implement centralized identity and policy enforcement for AI agents. Third, the regulatory landscape underlying Section 2.1, particularly the EU AI Act's phased enforcement, is still evolving, and specific compliance obligations should be verified against current regulatory text rather than this paper.

5.3 Future Work

The most valuable next step is empirical: instrumenting Pillar 5 audit trails across real enterprise agentic deployments to build the kind of longitudinal, cross-organization dataset that does not yet publicly exist, enabling genuine statistical validation of the risk-positioning and maturity claims sketched qualitatively here.

6. Conclusion

This paper synthesized the current standards landscape (NIST AI RMF, ISO/IEC 42001, the EU AI Act) and empirical security research on prompt injection, data exfiltration, and agentic misuse into a five-pillar Enterprise GenAI Governance Model (EGGM): identity and access management, prompt/output guardrails, policy enforcement, sandboxed tool orchestration, and continuous audit. Table 1 maps documented risk categories to the pillar responsible for mitigating them, and Section 4's illustrative tools give organizations a starting point for risk prioritization and maturity self-assessment. The framework's central claim is architectural, not statistical: governance controls concentrated at these five points can materially reduce agentic-AI risk exposure without requiring organizations to forgo the autonomy that makes agentic AI valuable in the first place. Rigorous empirical validation of this claim awaits the longitudinal incident data that Section 5.3 identifies as the field's most pressing research gap.

Conflict of Interest

The authors declare that they have no competing financial or personal interests that could have influenced the work reported in this paper.

References

Carlini, N., Tramer, F., Wallace, E., et al. (2021). Extracting training data from large language models. *Proceedings of the 30th USENIX Security Symposium*.

Cloud Security Alliance. (2024). *Agentic AI threat modeling framework: MAESTRO*. Cloud Security Alliance Research.

European Parliament & Council. (2024). *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*.

Google. (2023). *Google's Secure AI Framework (SAIF)*. Google Security Blog / Google Cloud documentation.

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISec '23)*.

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

Microsoft. (2024). *Planning red teaming for large language models and their applications*. Microsoft Learn — Responsible AI documentation.

MITRE Corporation. (2024). *MITRE ATLAS — Adversarial Threat Landscape for Artificial-Intelligence Systems*. https://atlas.mitre.org/

National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1).

National Institute of Standards and Technology. (2024). *Generative artificial intelligence profile* (NIST AI 600-1).

OWASP Foundation. (2025). *OWASP top 10 for large language model applications* (v2.0). https://owasp.org/www-project-top-10-for-large-language-model-applications/

Perez, F., & Ribeiro, I. (2022). Ignore previous prompt: Attack techniques for language models. *NeurIPS ML Safety Workshop*.

Wang, L., Ma, C., Feng, X., et al. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science, 18*(6), 186345.

Xi, Z., Chen, W., Guo, X., et al. (2023). The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*.

Zou, A., Wang, Z., Kolter, J. Z., & Fredrikson, M. (2023). Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.