International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 06, June 2026)

Agentic AI Governance: Preventing Agent Sprawl, Data Leakage, and Uncontrolled Automation

Anand Laxman Mhatre
Senior Program Manager / Technical Architect, Accenture, Artificial Intelligence
anand.mhatre@gmail.com

***Abstract*— Agentic AI is rapidly emerging as the next phase of enterprise artificial intelligence. Unlike traditional Generative AI assistants that primarily respond to prompts, agentic AI systems can plan, reason, use tools, execute tasks, and interact with enterprise applications. These capabilities create significant opportunities for productivity improvement, customer service transformation, workflow automation, and operational efficiency. However, as organizations begin deploying multiple AI agents across business units, the risks associated with agent sprawl, data leakage, excessive autonomy, and uncontrolled automation are increasing. Agentic AI governance is the discipline of defining policies, controls, accountability, monitoring, and risk management practices for AI agents operating within an enterprise. Without proper governance, agents may access sensitive data, execute unauthorized actions, duplicate capabilities, produce inconsistent outcomes, or create compliance risks. This paper explores the major governance challenges associated with agentic AI and proposes a structured governance model to help enterprises safely scale AI agents while protecting data, maintaining human oversight, and ensuring measurable business value.**

***Keywords*— Agentic AI, AI governance, agent sprawl, data leakage, uncontrolled automation, AI agents, Generative AI, enterprise automation, human-in-the-loop, AI risk management, prompt injection, access control, auditability.**

I. INTRODUCTION

Generative AI has already changed how organizations create content, summarize information, write code, support customers, and improve employee productivity. The next wave of AI adoption is focused on agentic AI, where systems are no longer limited to generating responses. Instead, they can pursue goals, break work into steps, use tools, interact with enterprise applications, and complete workflows with varying levels of autonomy. Gartner has predicted that more than 40 percent of agentic AI projects may be canceled by the end of 2027 due to escalating costs, unclear business value, or inadequate risk controls, making governance a critical success factor rather than a secondary concern [1].

This evolution creates a powerful opportunity for enterprises.

AI agents can assist contact center representatives, automate case creation, validate documents, generate summaries, monitor compliance, support IT operations, analyze data, and complete repetitive administrative work. Deloitte describes this shift as the emergence of a silicon-based workforce, but also emphasizes that organizations must redesign operations rather than simply place agents on top of existing inefficient processes [2].

However, the same capabilities that make agentic AI valuable also make it risky. When agents can retrieve information, call APIs, create records, update systems, trigger workflows, and interact with users, they must be governed like enterprise-grade digital workers. Each agent needs a defined purpose, owner, access boundary, risk classification, monitoring process, and escalation model. Without governance, enterprises may experience agent sprawl, data leakage, excessive autonomy, unclear accountability, and uncontrolled automation.

II. KEY INSIGHTS

The following insights summarize the key governance concerns and controls for enterprise adoption of agentic AI.

TABLE I: KEY INSIGHTS

Agentic AI introduces new enterprise risk: AI agents can take actions, use tools, retrieve data, and interact with systems, which creates higher risk than traditional chatbots.
Agent sprawl can reduce control and increase cost: Unmanaged agents across departments can duplicate work, increase spend, and produce inconsistent outcomes.
Data leakage is a major concern: Agents connected to enterprise knowledge bases, documents, APIs, and customer data must be controlled through access policies and data protection rules.
Excessive autonomy must be avoided: Not every workflow should be fully automated; high-risk decisions require human review and approval.
Governance must be built into the agent lifecycle: Controls should start from agent design and continue through development, deployment, monitoring, and retirement.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 06, June 2026)

Human accountability remains essential: Even when agents execute tasks, business owners must remain accountable for outcomes, exceptions, and compliance.

III. PROBLEM DEFINITION

Organizations are moving quickly to adopt agentic AI, but governance practices have not matured at the same pace. Many enterprises are still using governance models built for traditional applications, robotic process automation, or basic chatbots. These models are not sufficient for agentic AI systems that can make decisions, use external tools, and perform actions across workflows. NIST AI 600-1 provides a cross-sector profile for managing generative AI risk and can serve as a foundation for agentic AI controls, but enterprises must extend these principles to cover tool use, autonomy, agent lifecycle management, and human oversight [3].

A. Agent Sprawl

Agent sprawl occurs when multiple AI agents are created across an organization without centralized visibility, standards, ownership, or lifecycle management. One business unit may create an HR policy agent, another may create a finance reporting agent, and another may create a customer service agent. Over time, hundreds of agents may exist without a common registry or governance model.

This creates several problems. Agents may duplicate the same functionality, rely on outdated knowledge sources, use inconsistent prompts, follow different approval rules, or produce conflicting answers. It also becomes difficult to measure cost, performance, usage, and business value. Agent sprawl is similar to application sprawl or RPA bot sprawl, but it carries higher risk because agents can dynamically interpret instructions, retrieve information, and take action.

B. Data Leakage

Agentic AI systems often require access to enterprise documents, customer records, operational data, knowledge bases, CRM systems, ticketing platforms, emails, APIs, and reporting systems. If access is not properly controlled, agents may retrieve or expose sensitive information.

Data leakage can occur in several ways. An agent may summarize confidential information for an unauthorized user. A prompt injection attack may manipulate the agent into revealing system prompts or restricted data. A poorly designed retrieval system may return documents outside the user permission boundary. Logs may store sensitive information unnecessarily. Third-party tools may receive data that should remain internal.

OWASP identifies sensitive information disclosure, system prompt leakage, vector and embedding weaknesses, and excessive agency as major LLM application risks that are directly relevant to agentic systems [4].

C. Uncontrolled Automation

The value of agentic AI comes from its ability to act. However, uncontrolled action can create major enterprise risk. Agents may create tickets, update records, send communications, approve workflows, trigger financial transactions, modify code, or interact with external systems.

If agents are given broad permissions without approval rules, they may make incorrect changes at scale. For example, an agent could misclassify cases, send inaccurate responses, overwrite records, approve an invalid request, or trigger unnecessary downstream work. Uncontrolled automation is especially dangerous when agents operate without confidence thresholds, human-in-the-loop controls, rollback mechanisms, or exception handling.

D. Unclear Accountability

In traditional business processes, employees and application owners are accountable for decisions and outcomes. With agentic AI, accountability can become unclear. If an agent makes an incorrect decision, who owns the issue: the business team, IT team, model provider, platform team, data owner, or process owner?

Without a defined accountability model, organizations may struggle to investigate incidents, correct errors, respond to audits, and assign ownership for continuous improvement.

E. Security Vulnerabilities

AI agents introduce new attack surfaces. These include prompt injection, indirect prompt injection, insecure tool use, excessive permissions, model manipulation, system prompt leakage, insecure plugins, weak authentication, and untrusted data sources. Recent research on web-agent security demonstrates that agents can be vulnerable to prompt injection attacks when they interpret untrusted external content as instructions [5].

Because agents can interact with tools and systems, security threats can move beyond inaccurate responses and become operational incidents. A compromised agent may expose data, execute unauthorized actions, or manipulate workflows.

F. Lack of Observability

Many AI pilots do not capture enough operational telemetry. Enterprises may not know which agents are active, what data they accessed, what actions they performed, how often they escalated, or whether users accepted their outputs.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 06, June 2026)

Without observability, governance becomes reactive. Organizations need dashboards, logs, audit trails, performance metrics, risk alerts, and incident reporting to manage agentic AI at scale.

IV. AGENTIC AI GOVERNANCE

Agentic AI governance is a structured framework for managing AI agents across the enterprise. It defines how agents are proposed, approved, designed, built, tested, deployed, monitored, improved, and retired. The objective of governance is not to stop innovation. The objective is to enable safe innovation by giving teams clear standards, reusable patterns, and risk-based controls.

A strong governance model should include the following elements.

A. Agent Registry

Every enterprise AI agent should be registered in a central inventory. The registry should include the agent name, business purpose, owner, risk level, data sources, tools used, permissions, deployment environment, model version, approval status, and retirement date.

B. Risk Classification

Agents should be classified based on data sensitivity, action authority, user population, regulatory impact, business criticality, and level of autonomy.

C. Access Control

Agents should follow the principle of least privilege. They should only access the data, tools, and systems required to complete their approved function.

D. Human-in-the-Loop Controls

Human oversight is essential for high-risk workflows. Agents should escalate when confidence is low, when sensitive data is involved, when policy exceptions occur, or when an action has financial, legal, clinical, or compliance impact.

E. Policy Guardrails

Agents should operate within defined guardrails, including content filters, data loss prevention, restricted topics, tool-use limits, approval thresholds, prompt injection defenses, and prohibited actions.

F. Auditability

Every agent action should be traceable. Logs should capture user requests, retrieved sources, tool calls, system actions, approval steps, outputs, and exceptions.

G. Lifecycle Management

Each agent should have a lifecycle that includes design approval, testing, deployment, monitoring, periodic review, version updates, and retirement.

V. GOVERNANCE FRAMEWORK FOR PREVENTING AGENT SPRAWL

Preventing agent sprawl requires centralized visibility and federated execution. Enterprise standards should be centrally defined, while business units can still create agents within approved guardrails.

A. Establish an Enterprise Agent Review Board

An enterprise review board should evaluate proposed agents before they move into production. The board may include representatives from business, technology, security, privacy, compliance, legal, risk management, and enterprise architecture.

B. Create Standard Agent Design Patterns

Organizations should define reusable design patterns for common agent types, such as knowledge agents, workflow agents, customer service agents, compliance agents, reporting agents, and developer productivity agents.

C. Use a Central Agent Registry

The registry should act as the system of record for all approved agents. It should allow leadership to understand how many agents exist, who owns them, what they do, and whether they are still delivering value.

D. Define Ownership and Funding

Every agent should have a named business owner and technical owner. The business owner is accountable for process outcomes, while the technical owner is responsible for platform reliability, integration, and operational support.

E. Periodically Review Agent Value

Agents should be evaluated against business outcomes such as productivity improvement, cost reduction, quality improvement, cycle time reduction, user adoption, and compliance performance. Agents that do not provide value should be retired or redesigned.

VI. GOVERNANCE FRAMEWORK FOR PREVENTING DATA LEAKAGE

Data protection must be built into agentic AI from the beginning. Agents should not be allowed to freely search, retrieve, summarize, or transmit enterprise data without controls.

A. Enforce Role-Based Access

Agents should inherit the permissions of the user or process they support. If a user cannot access a document directly, the agent should not be able to retrieve or summarize it for that user.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 06, June 2026)

B. Apply Data Classification

Enterprise data should be classified by sensitivity, such as public, internal, confidential, restricted, regulated, or highly sensitive. Agents should be approved only for the data classes required by their use case.

C. Use Secure Retrieval-Augmented Generation

Retrieval-augmented generation should be connected only to approved knowledge sources. Retrieved content should respect document-level permissions, metadata, and retention rules.

D. Mask Sensitive Information

Agents should use data masking, redaction, and tokenization where appropriate. Sensitive data such as personally identifiable information, protected health information, financial data, credentials, and legal records should be protected from unnecessary exposure.

E. Monitor Prompts and Outputs

Prompt and output monitoring can help detect data leakage, policy violations, unusual access patterns, and attempts to extract restricted information.

F. Control Third-Party Tool Use

Agents should not send sensitive enterprise data to external tools unless the tool has been approved through security, privacy, legal, and procurement review.

VII. GOVERNANCE FRAMEWORK FOR PREVENTING UNCONTROLLED AUTOMATION

Automation controls are essential because agentic AI can perform real actions. Enterprises should define what agents are allowed to do, under what conditions, and with what approvals.

A. Define Action Boundaries

Agents should have clearly defined action limits. For example, an agent may be allowed to draft a response but not send it, recommend a case update but not approve it, or create a ticket but not close it without review.

B. Use Confidence Thresholds

Agents should act only when confidence is above an approved threshold. Low-confidence outputs should be routed to a human reviewer.

C. Require Approval for High-Risk Actions

Actions involving financial transactions, clinical decisions, legal commitments, customer notifications, production changes, or compliance determinations should require human approval.

D. Implement Kill Switches

Organizations should be able to quickly disable an agent if it behaves unexpectedly, creates operational risk, or becomes compromised.

E. Maintain Rollback Procedures

For agents that update systems, rollback procedures should be defined. This allows the organization to reverse incorrect actions and restore prior system states.

F. Continuously Test Agent Behavior

Agents should be tested against normal scenarios, exception scenarios, adversarial prompts, data boundary conditions, and tool-use failures. Prompt leakage research also shows the importance of continuously probing AI systems to detect whether internal instructions or proprietary configurations can be exposed [6].

VIII. RISK AND MITIGATION VIEW

TABLE II: RISK AND MITIGATION VIEW

Hallucinated or incorrect outputs: Use approved knowledge sources, retrieval-augmented generation, confidence scoring, and human review.
Unauthorized data access: Apply least privilege permissions, user-context access, data classification, and full audit logging.
Prompt injection: Test direct and indirect prompt attacks, separate instructions from data, restrict tool calls, and monitor unusual behavior.
Excessive agency: Limit agent permissions, require approvals for high-risk actions, and implement safe defaults.
Agent sprawl: Maintain an enterprise registry, ownership model, design standards, and periodic value reviews.
Unbounded consumption: Use rate limits, cost controls, quotas, and automated alerting for abnormal usage patterns.

IX. IMPACT OF AGENTIC AI GOVERNANCE

A. Improved Enterprise Control

Governance provides centralized visibility into all agents, their owners, purposes, data sources, and permissions. This helps leadership manage agentic AI as an enterprise capability rather than a collection of disconnected experiments.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 06, June 2026)

B. Reduced Security and Privacy Risk

Access controls, data classification, monitoring, and guardrails reduce the likelihood of data leakage and unauthorized system activity.

C. Better Business Value

Governance helps ensure that agents are aligned to measurable business outcomes. This reduces the risk of investing in agents that do not provide clear value.

D. Higher Trust and Adoption

Employees and stakeholders are more likely to trust AI agents when outputs are explainable, auditable, and supported by human oversight.

E. Stronger Compliance

Audit logs, human approvals, policy enforcement, and documentation help organizations respond to regulatory, legal, and internal audit requirements.

F. Safer Automation at Scale

Governance allows enterprises to scale AI agents responsibly by defining boundaries, escalation rules, and continuous monitoring practices. Emerging research also proposes using intelligent agents themselves as part of security monitoring and mitigation frameworks for LLM application risks [7].

X. CONCLUSION

Agentic AI has the potential to reshape enterprise work by enabling AI agents to plan, reason, use tools, and execute workflows. However, this power introduces new risks that traditional AI and IT governance models are not fully prepared to address.

Without governance, enterprises may face agent sprawl, data leakage, excessive autonomy, unclear accountability, security vulnerabilities, and uncontrolled automation. These risks can reduce business value and expose organizations to compliance and operational failures.

A mature agentic AI governance model should include an agent registry, risk classification, role-based access control, human-in-the-loop workflows, policy guardrails, auditability, lifecycle management, and continuous monitoring. Governance should be embedded from the design phase and continue through deployment and retirement.

Agentic AI should not be treated as a collection of isolated experiments. It should be managed as an enterprise digital workforce. Organizations that establish strong governance early will be better positioned to scale AI safely, improve productivity, protect sensitive data, and build trust in intelligent automation.

XI. SUMMARY

TABLE III: SUMMARY

Agent sprawl: Central agent registry and review board	Impact: Better visibility, reduced duplication, improved enterprise control
Data leakage: Role-based access, data classification, masking, and secure retrieval	Impact: Stronger privacy and information protection
Uncontrolled automation: Action boundaries, approval workflows, and confidence thresholds	Impact: Safer execution of enterprise workflows
Unclear ownership: Business owner and technical owner model	Impact: Better accountability and operational support
Security vulnerabilities: Prompt injection testing, tool-use restrictions, and monitoring	Impact: Reduced cyber and operational risk
Poor ROI: Value reviews and lifecycle management	Impact: Higher business alignment and measurable outcomes

References

[1] Gartner. (2025). Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027. Retrieved from https://www.gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027

[2] Deloitte Insights. (2026). Agentic AI Strategy: The Agentic Reality Check: Preparing for a Silicon-Based Workforce. Retrieved from https://www.deloitte.com/us/en/insights/topics/technology-management/tech-trends/2026/agentic-ai-strategy.html

[3] National Institute of Standards and Technology. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile - NIST AI 600-1. Retrieved from https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-generative-artificial-intelligence

[4] OWASP Foundation. (2025). OWASP Top 10 for Large Language Model Applications. Retrieved from https://owasp.org/www-project-top-10-for-large-language-model-applications/

[5] Evtimov, I., Zharmagambetov, A., Grattafiori, A., Guo, C., & Chaudhuri, K. (2025). WASP: Benchmarking Web Agent Security Against Prompt Injection Attacks. arXiv. Retrieved from https://arxiv.org/abs/2504.18575

[6] Sternak, T., Runje, D., Granosa, D., & Wang, C. (2025). Automating Prompt Leakage Attacks on Large Language Models Using Agentic Approach. arXiv. Retrieved from https://arxiv.org/abs/2502.12630

[7] Fasha, M., Rub, F. A., Matar, N., Sowan, B., & Al Khaldy, M. (2026). Mitigating the OWASP Top 10 For Large Language Models Applications using Intelligent Agents. arXiv. Retrieved from https://arxiv.org/abs/2601.18105