International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 5, Sep-Oct 2026
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Enterprise AI Agent Governance Framework

A Comprehensive Enterprise Architecture Approach for Governing Autonomous and Agentic AI Across SAP BTP, Business Data, Integration, Security, Risk, and Responsible AI

Sanjeeve Kumar Gajadi*
*(HCL Technologies, Bengaluru, Karnataka, India)
Email: sanjeevekumar.g@hcltech.com

****

Abstract:

Enterprise AI agents are evolving from conversational assistants into autonomous digital actors that can interpret goals, invoke tools, access enterprise data, orchestrate workflows, make recommendations, and execute business actions. This shift introduces a new governance challenge because conventional AI controls are often designed for isolated models rather than goal-directed agents that operate continuously across applications, APIs, identities, business processes, and external services. This paper proposes an Enterprise AI Agent Governance Framework that integrates enterprise architecture, responsible AI, identity and access management, business data governance, integration governance, security engineering, model risk management, human oversight, observability, auditability, and lifecycle controls. The framework defines governance layers for agent purpose, authority, identity, tools, data, memory, prompts, models, actions, approvals, runtime monitoring, incident response, and retirement. It also introduces policy-based decision gates, risk-tiering, action boundaries, multi-agent governance, evidence logging, human-in-the-loop controls, and measurable key performance and risk indicators. The proposed approach enables enterprises to scale agentic AI while preserving accountability, security, compliance, transparency, operational resilience, and business trust.

Keywords — Agentic AI, AI Agents, Enterprise AI Governance, Responsible AI, SAP BTP, SAP AI Core, SAP Joule, Identity Governance, Zero Trust, Enterprise Architecture, Human Oversight, AI Risk Management, AI Observability

****

I. INTRODUCTION

Enterprise AI agents represent a significant evolution in the use of Artificial Intelligence within digital enterprises. Unlike traditional analytical models that return a prediction or chatbot systems that primarily generate text, AI agents can interpret goals, plan tasks, select tools, retrieve contextual information, invoke APIs, call enterprise services, update records, coordinate with other agents, and execute multi-step workflows. This ability to act on behalf of users creates substantial opportunities for productivity, customer service, finance, procurement, supply chain, human resources, software engineering, cyber operations, and enterprise knowledge management.

As organizations move toward agentic automation, governance must shift from model-centric controls to end-to-end controls over autonomous behavior. The enterprise must govern not only the underlying model, but also the agent's identity, purpose, permissions, tools, memory, prompts, business context, data access, action boundaries, approval rules, inter-agent communication, and runtime behavior. A failure in any one of these layers may result in unauthorized transactions, data leakage, unsafe decisions, regulatory violations, financial loss, or reputational damage.

The proposed framework treats AI agents as governed enterprise digital actors. Each agent is assigned a business owner, technical owner, risk classification, machine identity, approved purpose, permitted tools, data domains, action scope, escalation path, lifecycle state, and auditable evidence trail. This provides a repeatable enterprise mechanism for scaling AI agents without losing accountability.

II. BUSINESS PROBLEM

Many organizations are piloting AI agents faster than their governance models are evolving. Business teams may deploy agents through low-code platforms, custom applications, vendor copilots, or autonomous workflow services while security, enterprise architecture, legal, compliance, and data governance teams remain disconnected from the design process. This creates fragmented controls and inconsistent approval practices.

Agentic systems introduce risks that are materially different from conventional software. An AI agent may dynamically choose a tool, generate a plan, adapt to new context, call another agent, or take an action that was not explicitly coded as a fixed workflow. Prompt injection, malicious content, over-permissioned service identities, memory poisoning, tool misuse, model

hallucination, policy bypass, chain-of-action errors, and cascading multi-agent failures can therefore become enterprise operational risks.

The core business problem is the absence of a unified governance model that balances autonomy with control. Enterprises need a framework that allows low-risk agents to operate efficiently while enforcing stronger human approvals, segregation of duties, security controls, and continuous monitoring for high-impact agents.

III. GOVERNANCE OBJECTIVES

The framework establishes the following governance objectives: business alignment, accountable ownership, least-privilege autonomy, responsible AI by design, data minimization, explainability proportional to risk, human oversight, secure tool use, continuous assurance, traceability, incident readiness, and lifecycle governance. These objectives transform AI agent governance into a measurable enterprise capability rather than an ad hoc review activity.

Governance decisions should be risk-based. A read-only knowledge assistant should not require the same control intensity as an autonomous finance agent that can approve payments or modify master data. The framework therefore classifies agents by business impact, data sensitivity, action authority, regulatory significance, and potential harm.

IV. ENTERPRISE ARCHITECTURE MODEL

The Enterprise Architecture model positions AI agent governance across Business Architecture, Application Architecture, Data Architecture, Integration Architecture, Security Architecture, Technology Architecture, and Governance Architecture. Business Architecture defines agent-enabled capabilities, value streams, ownership, decision rights, and business outcomes. Application Architecture identifies systems the agent can access and the transactions it may initiate. Data Architecture governs the information the agent may retrieve, transform, retain, or generate.

Integration Architecture governs tools, APIs, event channels, workflow services, message brokers, and external services used by the agent. Security Architecture governs authentication, authorization, machine identities, secrets, encryption, zero trust, runtime isolation, and threat detection. Technology Architecture governs model runtimes, vector services, agent orchestration, cloud platforms, and observability. Governance Architecture establishes policies, review boards, approval thresholds, risk acceptance, and audit requirements.

V. AGENT LIFECYCLE GOVERNANCE

Agent lifecycle governance begins with business justification and risk classification. Every proposed agent should have a documented use case, accountable business owner, intended users, expected benefits, data domains, action scope, risk tier, and success criteria before technical development begins. Architecture review then determines whether the use case requires an agent, a deterministic workflow, a conventional application, or another solution pattern.

During design and build, teams define the system prompt, model selection, tool registry, API contracts, memory strategy, data access, content filters, human approval points, fallback behavior, escalation handling, and observability requirements. Testing includes functional, adversarial, security, privacy, compliance, performance, and human-factor validation.

Deployment is permitted only after control evidence is complete. Runtime monitoring measures agent behavior, tool usage, failed actions, policy violations, human overrides, token and cloud consumption, anomalous requests, model drift, and business outcomes. Retirement removes machine identities, revokes credentials, archives evidence, deletes unnecessary memory, and updates architecture repositories.

VI. IDENTITY, ACCESS, AND AUTHORITY

An enterprise AI agent should never operate as an anonymous or shared technical process. Each production agent requires a distinct machine identity mapped to its business owner, runtime environment, and approved permissions. Identity governance should enforce least privilege, attribute-based controls, token lifetime management, credential rotation, privileged access management, and continuous access review.

Authority must be explicit. The framework distinguishes between observe, recommend, prepare, execute-with-approval, and autonomous-execution authority. A customer service agent may recommend a refund, while only an approved workflow may execute the financial transaction. High-impact actions should require policy evaluation and, where appropriate, human confirmation or dual control.

VII. TOOL AND API GOVERNANCE

Tools are the mechanism through which agents affect the enterprise. Every tool must therefore be registered, classified, versioned, secured, and governed. Tool metadata should describe purpose, owner, allowed operations, input constraints, data sensitivity, rate limits, authentication method, failure behavior, and audit requirements.

Agents should invoke tools through governed APIs rather than direct database access wherever possible. API gateways can enforce authentication, authorization, schema validation, throttling, threat protection, logging, and policy checks. Sensitive tools should expose narrow capabilities rather than broad administrative functions, reducing the blast radius of agent errors.

VIII. DATA, MEMORY, AND KNOWLEDGE GOVERNANCE

Agent quality depends on trusted enterprise context. Business data, documents, vector stores, knowledge graphs, and retrieval indexes must therefore be governed through ownership, metadata, lineage, retention, quality controls, and access policies. Retrieval-Augmented Generation should enforce authorization at retrieval time rather than relying only on application-level filtering.

Agent memory introduces additional risk because stored conversation context may contain personal, confidential, or regulated information. Memory should be purpose-limited, encrypted, retention-controlled, and logically separated by user or business context. Long-term memory should be justified by business need, with mechanisms for correction, deletion, and audit.

IX. RESPONSIBLE AI AND HUMAN OVERSIGHT

Responsible AI controls should include transparency, fairness, robustness, accountability, privacy, explainability, and human agency. The level of explanation required should correspond to the decision impact. Low-risk productivity agents may require basic disclosure, while agents influencing employment, credit, healthcare, legal, or safety-related outcomes require stronger evidence and human review.

Human oversight is not limited to a generic approval button. Effective oversight requires decision context, confidence indicators, source evidence, policy status, alternative recommendations, and the ability to reject or modify an action. Organizations should measure human override frequency and review whether overrides indicate a design weakness or policy gap.

X. SECURITY ARCHITECTURE

Security architecture for agentic AI should apply Zero Trust principles to users, agents, tools, APIs, models, data stores, and runtime services. Controls include identity federation, multi-factor authentication for human operators, machine identity, workload identity, mTLS, OAuth 2.0, secret vaulting, encryption, network segmentation, content inspection, secure sandboxing, and security event monitoring.

Agent-specific threats include prompt injection, indirect prompt injection through retrieved content, malicious tool descriptions, data exfiltration, privilege escalation, poisoned memory, compromised plugins, model abuse, and automated high-volume misuse. Defense requires layered controls rather than relying solely on prompt instructions.

XI. MULTI-AGENT GOVERNANCE

Multi-agent systems create new coordination risks because responsibility may be distributed across planner agents, specialist agents, validators, and action agents. The framework therefore requires clear role separation, message contracts, authority boundaries, traceable delegation, and escalation rules. An agent should not be able to transfer broader privileges to another agent than it possesses itself.

Coordination logs should record which agent initiated a task, which agent delegated it, what context was transferred, which tools were invoked, and which agent produced the final action. This enables root-cause analysis and accountability for complex agent chains.

XII. OBSERVABILITY AND CONTINUOUS ASSURANCE

Enterprise observability should capture agent requests, plans, tool calls, data sources, model versions, policy decisions, execution results, exceptions, human approvals, latency, cost, and business outcomes. Sensitive content may require masking or tokenization while preserving audit value.

Continuous assurance combines automated policy checks with periodic control testing. Organizations can define thresholds for unauthorized tool attempts, hallucination rate, human override rate, failed transactions, security alerts, data quality issues, and abnormal cost growth. Breaches should automatically trigger containment, reduced authority, or suspension of the agent.

XIII. RISK MANAGEMENT AND CONTROL TIERS

Risk tiering enables proportional governance. Tier 1 agents are informational and read-only. Tier 2 agents can create drafts or recommendations. Tier 3 agents can execute bounded business actions with approval. Tier 4 agents can execute high-impact or regulated actions and therefore require enhanced controls, independent validation, strong human oversight, and executive risk acceptance.

Risk assessments should evaluate business impact, legal obligations, data sensitivity, autonomy, reversibility, financial exposure, external connectivity, safety implications, and dependence on third-party models. Residual risk should be documented and reviewed through established enterprise risk governance.

XIV. IMPLEMENTATION ROADMAP

Implementation begins with governance principles and an enterprise inventory of existing agents. Organizations then define a common agent registration standard, machine identity model, risk taxonomy, tool registry, logging standard, approval workflow, and minimum security controls. Pilot implementations should focus on a small number of representative agent types across different risk tiers.

Subsequent phases introduce automated policy enforcement, centralized observability, control

dashboards, enterprise architecture integration, model and prompt registries, agent certification, and periodic recertification. The operating model should evolve through lessons learned from incidents, audit findings, regulatory changes, and new agent capabilities.

XV. KEY PERFORMANCE AND RISK INDICATORS

Key performance indicators should measure business value and governance effectiveness. Examples include agent task completion rate, cycle-time reduction, user adoption, productivity improvement, cost per successful task, approval turnaround time, and reuse of governed tools. Key risk indicators include unauthorized tool attempts, policy violations, human override rate, prompt injection detections, data leakage events, hallucination rate, failed high-impact actions, stale credentials, and unreviewed agent versions.

Executive dashboards should combine value and risk rather than optimizing only for automation volume. A highly automated agent that generates frequent exceptions may create more operational risk than value.

XVI. CONCLUSION

Enterprise AI agents can become a powerful digital workforce capability, but their autonomy changes the nature of enterprise risk. Governance must therefore extend beyond model documentation and include identity, authority, tools, data, memory, prompts, actions, multi-agent delegation, security, human oversight, observability, and lifecycle management.

The proposed Enterprise AI Agent Governance Framework provides a structured architecture for balancing innovation with accountability. By treating AI agents as governed enterprise actors and applying risk-based controls throughout their lifecycle, organizations can scale agentic AI while preserving trust, regulatory compliance, cybersecurity, operational resilience, and measurable business value.

TABLE I — ENTERPRISE AI AGENT GOVERNANCE CONTROL MODEL

Control Domain	Primary Control Objective	Example Evidence
Purpose & Ownership	Ensure accountable business use	Agent charter, owner, risk tier
Identity & Authority	Enforce least privilege	Machine identity, roles, approvals
Tools & APIs	Restrict executable capabilities	Tool registry, API policies
Data & Memory	Protect enterprise information	Lineage, retention, access logs
Responsible AI	Maintain trustworthy behavior	Bias, transparency, oversight tests
Security	Prevent abuse and compromise	Threat model, secrets controls, SIEM
Observability	Provide runtime evidence	Agent logs, tool traces, KPI/KRI dashboard
Lifecycle	Control change and retirement	Version approvals, recertification records

REFERENCES

[1] ISO/IEC 42001:2023, Information technology — Artificial intelligence — Management system.
[2] National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework (AI RMF 1.0), 2023.
[3] National Institute of Standards and Technology, AI RMF Generative Artificial Intelligence Profile, 2024.
[4] ISO/IEC 27001:2022, Information security, cybersecurity and privacy protection — Information security management systems.
[5] ISO 31000:2018, Risk management — Guidelines.
[6] The Open Group, TOGAF Standard, 10th Edition, 2022.
[7] ISACA, COBIT 2019 Framework: Governance and Management Objectives, 2019.
[8] AXELOS, ITIL Foundation: ITIL 4 Edition, 2019.
[9] Cloud Security Alliance, Security Guidance for Critical Areas of Focus in Cloud Computing.
[10] CNCF, Cloud Native Landscape and Kubernetes Documentation.
[11] SAP SE, SAP Business Technology Platform Documentation.
[12] SAP SE, SAP AI Core and SAP AI Launchpad Documentation.
[13] SAP SE, SAP Integration Suite Documentation.
[14] SAP SE, SAP HANA Cloud Documentation.
[15] SAP SE, SAP Cloud Identity Services Documentation.
[16] SAP SE, SAP Cloud ALM Documentation.
[17] SAP SE, SAP Business Data Cloud and SAP Datasphere Documentation.
[18] SAP SE, SAP LeanIX Enterprise Architecture Documentation.
[19] SAP SE, SAP Signavio Process Transformation Suite Documentation.
[20] OECD, Recommendation of the Council on Artificial Intelligence.
[21] European Union, Artificial Intelligence Act.
[22] IEEE, Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems.
[23] Project Management Institute, PMBOK Guide, Seventh Edition, 2021.
[24] DAMA International, DAMA-DMBOK2: Data Management Body of Knowledge, 2017.

AUTHOR BIOGRAPHY

Sanjeeve Kumar Gajadi is an SAP Enterprise Architect and Lead Consultant with extensive experience in enterprise architecture, SAP Business Technology Platform, data and analytics, integration, cloud architecture, security, governance, and AI-enabled transformation. His professional interests include agentic AI governance, responsible AI,

enterprise architecture, cloud platforms, security, and intelligent automation.

ACKNOWLEDGMENT

The author acknowledges the contributions of standards bodies, enterprise architecture communities, SAP documentation teams, and responsible AI practitioners whose published guidance has informed the concepts synthesized in this paper.

