THE ARCHON

TECHNOLOGY INTELLIGENCE

EXECUTIVE BRIEF

AI agents are everywhere—who decides their actions?

Enterprise adoption of autonomous AI agents is rapidly growing, with many organizations now deploying at least one such agent.

→ Agents power customer service, code generation, and decision support.

→ Rapid diffusion outpaces formal governance structures.

CONTROL FRAMEWORK

Regulatory landscape is converging

Governments are issuing binding rules that define who must control AI agents and how they are evaluated for safety, fairness, and transparency.

→ U.S. Executive Order 14110 (Oct 2023) mandates risk-based governance for AI systems.

→ EU AI Act (effective 2025) classifies high-risk agents and sets conformity obligations.

→ U.S. AI Bill of Rights outlines user-centric safeguards for AI-driven decisions.

CONTROL FRAMEWORK

Emerging governance frameworks

Organizations are creating structured approaches to assign control, assess risk, and document compliance for AI agents.

- → A lifecycle-wide view of risk guides development and deployment decisions.
- → Defined requirements outline how governance systems should function.
- → Core principles emphasize transparency, accountability, and human-centric values.

CONTROL FRAMEWORK

Ownership models shape control

Who owns the model, data, and runtime environment determines the locus of responsibility for AI agent behavior.

→ Enterprise-built agents (e.g., internal fraud-detection models) – full internal control.

→ Platform-hosted agents (e.g., Microsoft Copilot, OpenAI API) – provider retains model updates and policy enforcement.

→ Open-source community agents (e.g., Hugging Face models) – shared stewardship, variable governance.

CONTROL FRAMEWORK

Technical control mechanisms

Effective governance combines policy, provenance, and runtime safeguards to keep AI agents aligned with organizational intent.

- → Policy enforcement via model-level guardrails (Google Vertex AI Guardrails).
- → Runtime monitoring and anomaly detection (Microsoft Responsible AI Toolbox).
- → Model provenance registries to track version, data lineage, and ownership.

CONTROL FRAMEWORK

Risk, compliance, and liability

Unclear control creates gaps in oversight, leading to penalties and potential civil liability for outcomes driven by AI.

→ Guidance warns against deceptive or unfair automated decisions.

→ Non-compliant high-risk AI can result in significant penalties.

→ Audit trails and explainability are needed for privacy and fairness assessments.

EXECUTIVE TAKEAWAY

Executive takeaway

CIOs, IT Directors, and CISOs need to set clear ownership, embed enforceable policies, and continuously monitor AI agents to reduce risk and ensure trustworthy operations.

→ Define who owns the model, data, and runtime environment.

→ Adopt a governance framework based on recognized best practices.

→ Deploy guardrails and provenance tools for real-time compliance.