anjuna

WHITEPAPER

Why runtime supervision is essential for secure, scalable, and cost-controlled enterprise agentic AI

Deploying enterprise AI agents demands a new security model

Agentic AI is rapidly moving from experimentation to operational deployment. Unlike traditional generative AI systems that answer questions or generate content, agentic systems can plan, decide, invoke tools, access enterprise data, and execute multistep actions with limited or no human intervention. That shift creates a fundamentally different enterprise risk profile. The Open Worldwide Application Security Project (OWASP)'s Top 10 for Agentic Applications frames agentic systems as a distinct security category, with risks such as goal hijacking, tool misuse, identity and privilege abuse, memory poisoning, and multi-agent failure modes. Beyond the security risks, agents come with serious economic and business risks including token budget exhaustion, unexpected behavior, and requesting high-cost cloud resources (e.g. high-volume data retrieval) necessary to meet goals.

This is important to note because while agentic AI can bring business cost efficiency, the related security and business risks can also create financial loss that impacts an enterprise's bottomline. In fact, Gartner predicts that more than 40 percent of agentic AI projects will be canceled by the end of 2027 because of escalating costs, unclear business value, or inadequate risk controls.

For enterprise leaders, the implication is straightforward: an AI agent cannot be allowed to run without governance. It must be supervised by a model that places policy, control, observability, and intervention above the agent itself. It constrains what the agent can access, governs how it can act, creates accountability for every sensitive action, and limits both breach blast radius and runaway cost. This replicates a proven hierarchy in the human context - mentoring, leadership and supervision with integrity and trust combined with policy to stay on track and within acceptable measurable risk controls.

The enterprises that succeed with their agentic AI journey will deploy agents that are bounded, provable, observable, and governed at runtime.

The shift from generative AI to agentic AI changes the risk profile

The first wave of enterprise AI centered on providing assistance in drafting, summarizing, translating, and accelerating human work. Recent AI models change the way they work. Now, agentic AI can reason through tasks, call APIs, chain actions, coordinate with other tools, and operate with delegated authority.

That sounds like productivity, and it is. BCG says effective AI agents can accelerate business processes by 30 percent to 50 percent, but it also stresses that organizations must balance AI autonomy with human oversight and embed controls from day one. This shift means that errors are no longer confined to bad outputs that are still handled by humans. In agentic systems, the consequence of failure can be immediate and damaging: data exposure, improper approvals, policy violation, unintended transactions, lateral movement across systems, or uncontrolled spend across models, tools, and workflows. As "digital employees," agents are now a new type of insider threat. They are vulnerable to new hacks and manipulation. Together with their own potentially dangerous intelligent capabilities, they are simultaneously powerful and yet exposed to new risks to the enterprise.

As AI systems gain agency, their risk profile starts to resemble a combination of application security risk, identity risk, operational risk, and insider risk.

The OWASP's agentic application framework highlights this point: autonomous systems face risks beyond classic prompt injection alone. Goal hijacking can redirect the mission of the agent. Tool misuse can convert legitimate enterprise integrations into attack paths. Identity and privilege abuse can allow an agent or attacker to act beyond intended authority. Supply chain and memory vulnerabilities can influence behavior over time in ways that are difficult to detect.

Another perspective: rather than treating AI risk as a narrow model problem, NIST frames it as a lifecycle risk management challenge that spans design, development, deployment, operation, and oversight.

For enterprises, these risks show up in three especially important ways.

First, the attack surface expands dramatically. Agents can be granted access to sensitive tools, systems, and data stores. Agents, as software systems, can be attacked, and they can actively leak and disperse data autonomously. If compromised, manipulated, or over-permissioned, they can become highly efficient pathways for data theft or operational disruption. The result can be damaging very quickly for today's enterprise that operates across different environments. A hybrid environment's interconnected cloud and on-premises systems mean a compromised AI agent doesn't stay contained. Its legitimate access becomes the attack path, opening up the attack surface and accelerating the blast radius across the entire enterprise.

Second, accountability becomes harder to track. Without strong supervision and auditability, it can be difficult to determine why an agent acted the way it did, which data informed it, which tools it used, what policy it should have followed, and whether the failure was malicious, accidental, or systemic. In the event of a breach and without knowing why an agent acted the way it did, enterprises are left unable to assign accountability, close the gap that enabled the failure, or defend their decisions to regulators and stakeholders.

Third, costs grow unpredictable. Without clear boundaries, agents can multiply model calls, recurse through workflows, overconsume external services, and introduce operational overhead that's difficult to detect or attribute. Unchecked agent proliferation can trigger cascading cost overruns, resulting in spiraling costs

New risks require new security approach

The answer to enterprise agentic AI security risk is not to prevent deployment. In fact, Anjuna believes the right approach is to manage the agents using an independent control layer to benefit from the improved productivity without introducing vulnerabilities to the enterprise.

This supervisory approach means that trust does not reside in the agent alone. Instead, the independent control layer governs what the agent is allowed to do, under what conditions, with which data, through which tools, and with what level of evidence and accountability.

Common industry practices agree with this built-in, automated approach. KPMG's guidance for the agentic era emphasizes the need for default scope boundaries, oversight requirements, immutable logging, and fail-safe protocols. McKinsey recommends explicit oversight mechanisms, monitoring, anomaly detection, escalation triggers, and governance across the full lifecycle.

This control layer should include the following features:

- **Policy-based runtime control:** The enterprise should define according to their governance frameworks what the agent may access, which tools it may use, which actions are permitted, and what conditions require review or denial.
- **Least-privilege identity and access:** Agent identities should be scoped to purpose, time, and policy - and by context. Access should be bounded and revocable, not broad and persistent. An agent with even short-term access to a resource in one moment should not necessarily have access to the same resource moments later unless context and intent is valid.
- **Live mediation of tool use and data access:** Every call to a tool or sensitive dataset should be subject to inspection, policy evaluation, and, where appropriate, approval, including approval conditioned on limiting data use to what is strictly required to meet stated goals.
- **Real-time continuous monitoring and intervention:** The enterprise must be able to observe actions in real time, detect drift or anomaly, pause or stop execution, revoke access, or force fallback to safer modes.
- **Immutable audit and traceability:** Sensitive operations should produce durable evidence: what the agent attempted, what it was permitted to do, which policies applied, what data was touched, and what outcome followed.
- **Incorruptible operations with strong evidence and proof:** As agents grow more capable with advances in new models, the supervisory function must itself be held to the highest integrity standards. A system that guides operations and governs data access must ensure those oversight functions are both tamper-proof and private.

The value of trusted enterprise agentic AI

Managing and controlling enterprise agentic AI is not only a security issue; it is also a business issue.

A context-aware, intelligent, and incorruptible supervisory approach reduces **risk of breach** as well as the **attack surface** by narrowing the actions an AI agent can take and limiting the blast radius when something goes wrong. For example, an agent may need data to figure out a problem. This requires access, data and tool use. However, the tool may furnish information that contains more data than the agent needs to meet the goal. The information may itself represent compliance or leakage risk (for example, PII, customer data, or other sensitive response information).

By equipping AI agents with only what they need and precisely when they need it, the supervisory system enforces zero-trust assessments across pending, prior, and goal-aligned access requests. With full visibility into agent intent and strict limits on unnecessary data exposure, agents remain on track to meet objectives, while security teams retain complete audit proof, measured trust, and integrity guarantees. This approach goes far beyond the basic block and tackle security techniques that work in deterministic systems, but don't align to an intelligent and non-deterministic agent working for the enterprise with AI-powered reasoning.

It improves **economic discipline** by enforcing boundaries around model consumption, tool invocation, retries, workflow depth, and downstream service usage. That matters because many enterprises are discovering that the value of AI disappears quickly when operational spend is not controlled.

It strengthens **regulatory and governance readiness** by making oversight demonstrable. Regulatory gaps can be detected instantly, while a complete record of every agent decision is available for audit purposes, helping the enterprise stay on compliance track.

Finally, it accelerates **safe adoption at scale** because security, compliance, and business teams can align around one operating model. That gives enterprises a way to leverage the productivity of agentic AI and say yes to innovation without surrendering control.

The Path Forward

The question is no longer whether agentic AI will become mainstream deployment in the enterprise. It has already begun. The question is whether organizations will deploy it with the control architecture required to make it safe, scalable, and economically defensible.

At Anjuna, we believe that the future belongs to enterprises that combine AI autonomy with provable control. Agents are powerful but they should not operate as implicitly trusted staff. They need a control layer that enforces policy, mediates access to tools and data, limits agency to approved boundaries, and creates evidence of control for security and compliance teams.

This is how enterprises move from AI experimentation to trusted execution. It is how they reduce attack surface while containing cost. And it is how agentic systems deliver real business value in production, rather than becoming another source of unmanaged risk.

With this level of control, enterprises can govern access, contain cost, prove accountability, and adopt the next phase of agentic AI deployment with confidence.

About Anjuna Security

Anjuna gives enterprises the freedom to govern and scale AI agents across different environments with confidence. Built on confidential computing technology, Anjuna's platform provides the central control to ensure AI agents act only as intended, stay within budget, and continuously comply with regulatory frameworks. With Anjuna, enterprises from regulated industries including financial services, government, defense, and healthcare can move fast with trusted autonomous AI, efficiently achieving goals without increasing risk and losing security oversight.

For more information, visit **anjuna.io**