Market Guide for Guardian Agents

25 February 2026 - ID G00836388 - 35 min read

By: Avivah Litan, Daryl Plummer, Carlton Sapp, Dionisio Zumerle, Tom Coshow, Max Goss, Lauren Kornutick

Initiatives: Delivery of Functional Responsibilities; Govern Agentic AI for Greater Business Autonomy

Guardian agents supervise AI agents, helping ensure agent actions align with goals and boundaries. They monitor and block risky actions and are evolving from a collection of services to autonomous agents that enforce policies across platforms. AI leaders can use this Guide to understand the market and vendors.

Overview

Key Findings

- AI agents introduce new risks that outpace human review, yet most enterprises are unprepared to manage them due to fragmented organizational structures and ongoing challenges with discovery.
- Vendors providing AI agent development and other IT and security platforms are incorporating guardian agent (GA) capabilities that users can leverage, while dozens of startups and specialized providers offer GA capabilities that address domain-specific issues.
- Organizations need independent GAs to manage their agents across clouds and hosting environments, information repositories, and multiple identity systems, and to enforce their own specific acceptable use policies.
- GA technology is still nascent, focused on observation and posture management with very limited in-line blocking or remediation but is evolving quickly. Integration across AI agent platforms is challenging due to a lack of standard interfaces.

Recommendations

- Launch a cross-functional initiative to systematically discover, inventory, map and manage all AI agents — sanctioned and unsanctioned — across the organization. Leverage AI discovery solutions provided by best-in-class guardian agent vendors.
- Trial emerging guardian agents now to gain early expertise in safely overseeing increasingly autonomous AI systems, securing a lasting competitive advantage as these tools evolve into mature, full-scale automated AI agent overseers.
- Invest in emerging GA oversight mechanisms that aid in continuous AI agent discovery, access management, assurance, monitoring, and improvement.
- Prioritize GA solutions independent of AI agent platforms to ensure cross-cloud governance, full enterprise information governance and avoid vendor lock-in. Independent solutions should integrate with and complement GA solutions embedded in AI agent platforms for optimal coverage and results.
- Implement metagovernance controls for guardian agents themselves to mitigate their own risks of deviant or destructive actions and behavior.

Strategic Planning Assumptions

By 2029, independent guardian agents will eliminate the need for almost half of incumbent risk and security systems protecting AI agent activities in over 70% of organizations.

By 2027, over 70% of AI agent identity providers will classify the sensitivity of data agents interact with as part of granting and scoping access rights.

Through 2028, at least 80% of unauthorized AI agent transactions will be caused by internal violations of enterprise policies concerning information oversharing, unacceptable use or misguided AI behavior rather than from malicious attacks.

Market Definition

Gartner defines guardian agents as a blend of AI governance and AI runtime controls in the AI TRiSM framework that supports automated, trustworthy and secure AI agent activities and outcomes. Guardian agents use AI-based and deterministic evaluations to oversee AI agents and their interactions with tools, data, APIs and humans. They are evolving from a collection of human-directed automated oversight services into semiautonomous or fully autonomous agents capable of formulating and executing action plans, and redirecting or blocking actions to align with intended agent goals.

Enterprise adoption of AI agents is accelerating, outpacing maturity of governance policy controls. As AI agents become more autonomous and embedded in critical workflows, the risks of operational failure and noncompliance escalate.

Guardian agents:

- Provide advanced oversight services that supervise and support real-time, automated and adaptive assurance of AI agents used at an organization.
- Give the AI agents they supervise situational awareness and intelligence needed to act within the scope of the agent's preset intentions.
- Continuously review, observe or adjust agent activities and outputs for acceptable use, data leakage, third-party risks, safety, security and compliance — enabling human or AI follow-up or automated interventions when necessary.
- Coexist with other guardian agents and integrate to blend capabilities for optimal outcomes. For example, they operate in domains such as agent identity and access management and information governance where integrated decisions are imperative.

Most AI agent platform vendors provide their own guardian agent capabilities, but independent enterprise-owned guardian agents are also required to support:

- **Cross-cloud and hosted environments.** No cloud provider can unilaterally enforce runtime control over AI agents once they operate or delegate across another provider's cloud or on-premises environments.
  - Vendor safeguards and controls typically stop at their own cloud borders.
  - Cross-cloud policy extensions currently rely on opt-in partnerships and SDKs.
  - These opt-in mechanisms provide limited, voluntary cross-cloud governance through explicit registration and shared protocols.
  - Without such opt-ins, cross-cloud agent interactions remain completely ungoverned. No single provider can close this governance gap on its own.
- **Cross-platform identity and access management.** Enterprises need to identify all agents regardless of registries or methods they use when created.

- **Cross-platform information governance.** Enterprises need to interface with and protect all information across multiple platforms and environments.

Only a neutral, trusted guardian agent layer with multiple guardian agents performing separate but integrated oversight functions can enforce routing across all providers. Thus, the guardian agent acts as the missing universal enforcement mechanism.

Mandatory Features

AI visibility and traceability must be underpinned by the following features:

- **The AI agent catalog** inventories all agents — registered, unregistered, official, custom, third-party, shadow or rogue — within the organization's network. It scores risks and tracks them over time. It stores metadata (agent cards) detailing identity, capabilities, interaction endpoints (e.g., APIs, gateways), authentication requirements and other details to enable discovery, interoperability and secure collaboration between agents, regardless of platform.
- **Maps:** Visual or structured maps show how AI agents integrate with human workflows, systems, tools, and other agents throughout their life cycle. These maps highlight connections, data flows, risks and dependencies, supporting governance, compliance and management of agent sprawl.
- **Ownership mapping:** Ownership mapping tracks the human and machine owner of each AI agent and its artifacts, attributing responsibility for design, maintenance and outputs. It captures full lineage from creation to deployment, enabling accountability, compliance auditing and governance controls.
- **Audit trails:** Comprehensive, tamper-evident logs record every change, interaction and decision involving AI agents and their artifacts. Audit trails enable forensic analysis, regulatory compliance, accountability and detection of unauthorized modifications or incidents.

Continuous assurances and evaluation must be underpinned by the following features:

- **AI agent posture management:** Posture management aggregates life cycle metadata to provide real-time awareness of agents' security, compliance and operational health. It integrates with inspection tools for dynamic enforcement, enabling continuous risk management and anomaly remediation.

Runtime inspection and enforcement must be underpinned by the following features:

- **Agent alignment:** Agent alignment ensures that AI agents' actions and outputs match defined intentions, goals and governance policies, preventing unintended behaviors. Continuous evaluation raises flags or intervenes when deviations occur, supporting trust and safe scaling.
- **Anomaly detection:** Anomaly detection flags suspicious or unusual AI agent activities, such as abnormal tool use or behavioral shifts, using rule-based or machine learning methods. High-confidence anomalies trigger autoblocking and alerts, helping prevent harm and catch threats before impact.
- **Runtime adaptation:** Dynamically fuse real-time threat intelligence, internal data changes and external signals into enriched contextual feeds for proactive runtime detection and adaptive automated responses.

Common Features

AI visibility and traceability commonly features:

- **AI agent identity discovery:** Automated discovery scans all environments to identify AI agents, their digital identities, credentials, and access rights. This process reveals unmanaged agents, orphaned credentials, and excessive privileges, supporting least-privilege enforcement and compliance with AI governance standards.
- **AI agent data mapping and lineage:** Data sources, pipelines and flows connected to AI agents are cataloged and visually presented while tracking data access and transformations over time. This enables enforcement of data policies, detection of risky usage and integration with governance tools, and supports regulatory audit trails.

Continuous assurances and evaluation commonly features:

- **AI agent security testing:** Security testing identifies and mitigates vulnerabilities in AI agents and their ecosystems through automated scans, red teaming and behavioral fuzzing (i.e., stress-testing with random inputs). These measures protect against attacks like prompt injection, tool misuse and unsafe outputs, ensuring more trustworthy deployments.

- **Risk and control validation for AI agents:** Regular assessments check for output bias, data leakage, misalignment and risky behaviors, validating adherence to risk thresholds and control objectives. This helps detect drift or shadow activities that could harm the organization.
- **AI agent compliance reporting:** Generates reports to demonstrate AI agents' adherence to regulatory standards, risk frameworks and organizational policies, covering inventory, risk, lineage and controls. Automated reporting supports audits, governance, and executive oversight while reducing manual effort.

Runtime inspection and enforcement commonly features:

- **Automatic blocking:** Real-time protection mechanisms detect and immediately block high-risk or anomalous behaviors in AI agents, such as unauthorized actions or privilege escalation. Advanced analytics and policy guardrails enable autoblocking and minimize operational disruption, providing proactive threat prevention.
- **Autoremediation:** Automatically applies corrective actions — like revoking privileges or quarantining agents — in response to detected anomalies or security events. The system refines these responses over time, reducing manual intervention and improving resilience.
- **Continuous compliance monitoring:** maps AI agent activities and data flows against regulations and policies, detecting violations in real time. The system can trigger alerts, enforce controls, and generate audit-ready logs, ensuring regulatory defensibility and proactive risk management.

What the guardian agent Market Guide does not Include:

This Market Guide excludes providers who do not provide native guardian agent controls in all three mandatory categories (see Market Definition for elaboration):

- AI visibility and traceability
- Continuous assurances and evaluation
- Runtime inspection and enforcement

Market Description

The guardian agent market is in its early stages, with most tools still under development and in limited deployment. As noted in the market definition, guardian agents are evolving from a collection of human-directed automated AI agent oversight services into semiautonomous or fully autonomous agents.

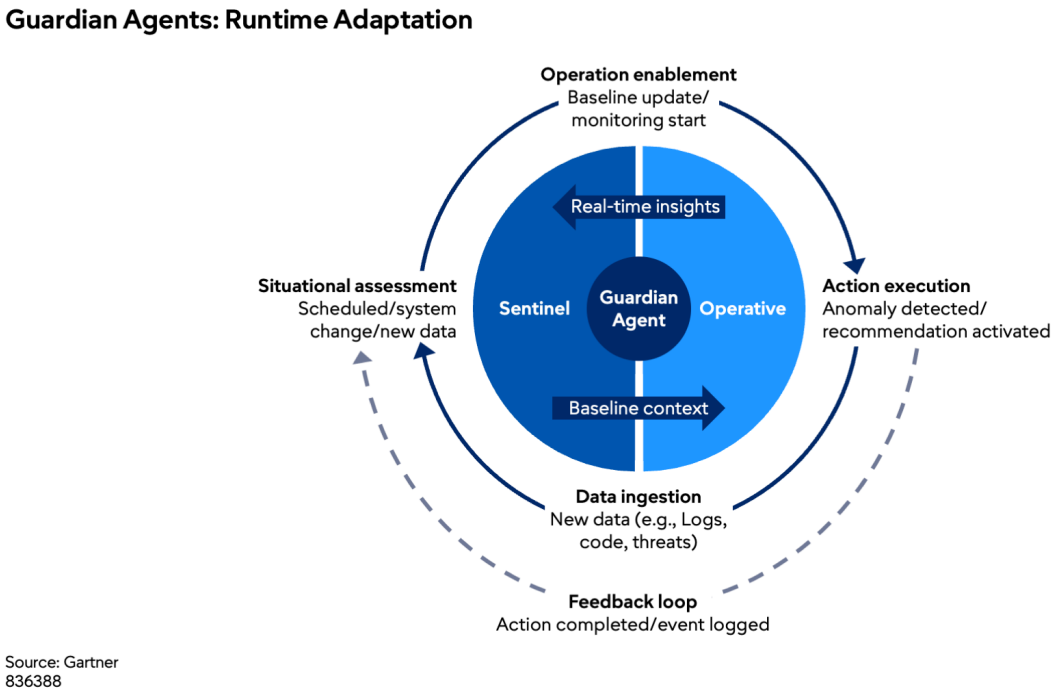
Although the market is very early, vendor solutions are advancing quickly to meet organizational and scalability demands. Through 2027, multiple guardian agents that specialize in different domains and tasks will coexist and integrate to assure and secure AI agent behavior and performance for an enterprise.

Current Market Landscape

Today, guardian agent deployments are mainly prototypes or pilots, although advanced organizations are already using early versions of them to supervise AI agents. Most guardian agent tools today support passive monitoring using observability and evaluation gateways — to provide visibility into agent activities, with limited real-time intervention and remediation. Fully autonomous gauardian agents capable of enforcing policies or corrective actions in real time are mostly confined to research and proof-of-concept efforts. Nonetheless, innovation is accelerating as enterprises pursue scalable, automated oversight, using solutions that match the rapid pace of AI agent actions.

See Figure 1 below for a depiction of emerging guardian agents' functionality as semi- or fully autonomous agents, combining posture assessments with runtime intervention and remediation, and see Note 8 for an outline of evaluation methods guardian agents use. Note that Sentinels provide the environmental context, posture assessment and situational awareness needed by Operatives at runtime to identify risks and threats and prioritize responses.

Figure 1: Guardian Agents: Runtime Adaptation

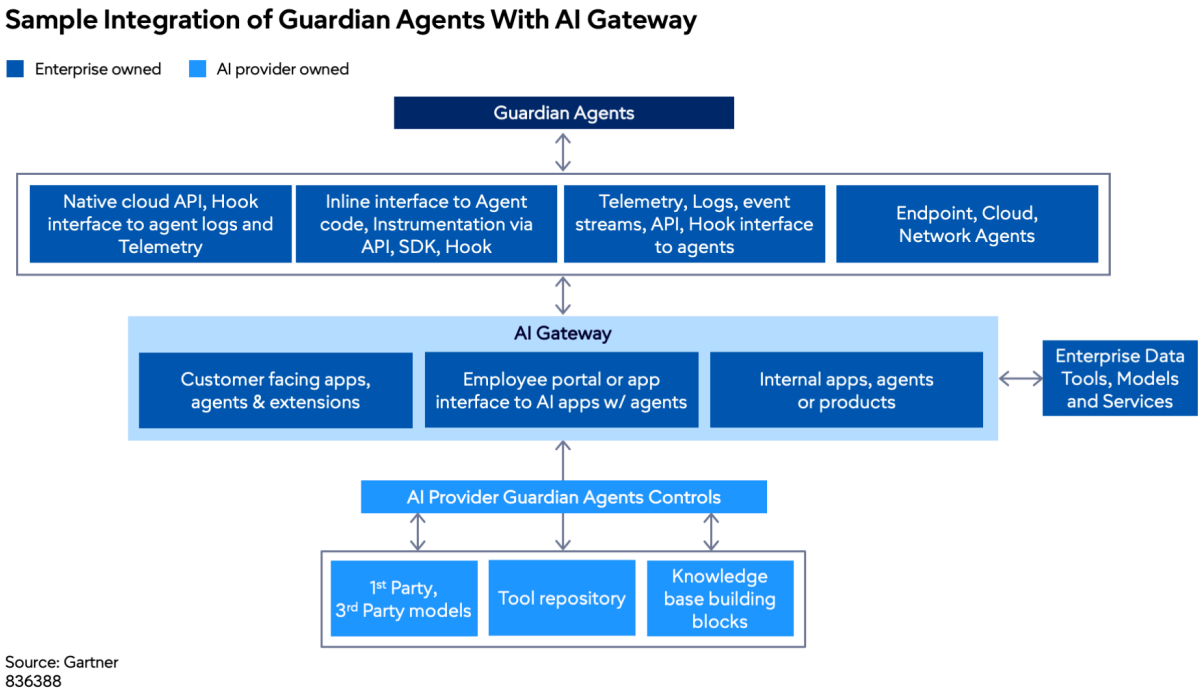


Emerging Delivery and Integration Models

Multiple delivery models are still maturing, and many are still experimental, as vendor offerings expand (see Market Definition for more information on delivery models): Figure 2 gives an example of guardian agent integration using

- AI/MCP gateways: Centralized systems to monitor and enforce policies on agent traffic
- Embedded/In-line runtime modules: Observability and policy modules within agent platforms, such as AI agent management platforms (see Note 2), or LLM proxies
- Stand-alone oversight platforms: Tools for aggregating and analyzing agent logs
- Orchestration layer extensions: Plugins for multiagent workflow oversight
- Hybrid edge-cloud model: distributed oversight across edge hardware and remote cloud-based resources (becoming more important as AI agent activities and ecosystems become endpoint-centric)
- Coordination mechanisms: Standards, APIs, and Hooks for unified oversight and policy enforcement

Figure 2: Sample Integration of Guardian Agents With AI Gateway



DIY Approaches

Organizations with advanced technical capabilities can use open-source frameworks, adapt monitoring tools, build custom policy engines, and apply model-agnostic wrappers for agent oversight. These approaches require significant expertise and resources, making them suitable for organizations with strong in-house teams. (See Note 3 on DIY tools.)

Vendor Landscape

The market features a range of vendors that serve different enterprise buyers and AI TRiSM (trust risk and security management) controls. Many of these vendor segments intersect or overlap, for example “AI agent development and governance platforms” sometimes partner with AI Risk and Security specialists to enhance their own guardian agent functions. Agent Identity vendors often partner with Information Governance vendors to provide data classification context to better inform agent identity and permissioning decisions. See Market Analysis section below for analysis on vendor categories for guardian agents.

Enterprises Require Independent Guardian Agents

A neutral, trusted guardian agent layer with multiple guardian agents performing separate but integrated oversight functions enforces routing across all providers. Thus, the guardian agent acts as the missing universal enforcement mechanism.

Most AI agent platform vendors provide their own guardian agent capabilities, but independent enterprise-owned guardian agents are also required to support:

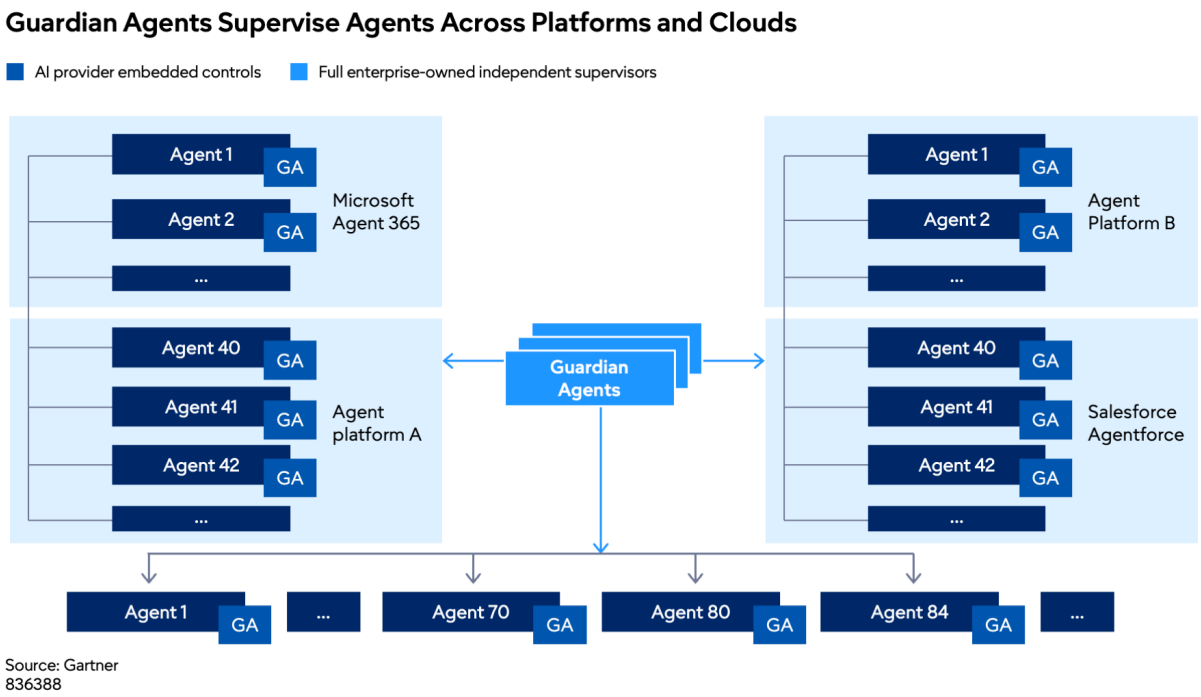
- **Cross-cloud and hosted environments.**
- **Cross-platform identity and access management.**
- **Cross-platform information governance.**

(See linked Market Definition for more details on this architecture requirement which is also depicted in Figure 3.)

Guardian agent packaged capabilities are what actually matters in AI agent systems — not the platform that hosts them — as these lightweight, high-performance agents can be embedded into any system, making small, agile capabilities the defining architecture and system of the future.

Figure 3 shows how independent guardian agents complement the first-party guardian agents embedded in AI agent platforms.

Figure 3: Guardian Agents Supervise Agents Across Platforms and Clouds



Market Direction

Guardian agent spend: By 2028, organizations will allocate between 5% and 7% of total agentic AI spend to guardian agents, up from less than 1% today — including embedded platform solutions and independent oversight layers — as enterprises prioritize governance, security, risk management, and compliance in increasingly complex multiagent environments.

Platform expansion and multiagent ecosystem governance: AI agent development and governance platforms are extending capabilities to support local, opt-in guardian agent governance and multicloud, multiagent system interoperability. These advancements facilitate the integration of governance controls across hybrid deployments and accelerate enterprise adoption of AI agents but depend on opt-in agreements from hosting systems in foreign environments.

Near-term market consolidation and platform unification: The guardian Agent market will undergo rapid consolidation as large security and network vendors continue to acquire AI TRiSM-focused startups with GA capabilities, and embed their controls into unified platforms. This consolidation will reduce ecosystem fragmentation and enable acquiring vendors to capture a larger share of market economics. Recent acquisitions, such as Palo Alto Networks' purchase of Protect AI and Check Point's acquisition of Lakera in 2025, exemplify this trend toward integrating specialized governance tools into broader security platforms. However, this trend will be overshadowed by independent guardian agents that surpass existing platforms in GA functionality and eventually supersede the need for many security systems used to protect AI agents today. The platforms will try hard to acquire the best-in-class independent guardian agents, but a few independents will remain intact to gain dominant GA market share.

Enterprise adoption of independent guardian agents: Organizations will increasingly deploy enterprise-owned guardian agent layers atop embedded platform tools. These independent agents, which will have specific domain expertise, will traverse multicloud, IAM, and data environments, providing unified oversight and mitigating risks of silos and vendor lock-in. This trend supports the push toward "verified accountable autonomy" in multiagent systems, where independent supervisors enforce rules and generate audit trails to prevent rogue behavior across distributed environments.

Guards for the guardians: As enterprises deploy guardian agents, it becomes essential to implement robust metagovernance controls to prevent misalignment, security breaches, and operational risks from the guardian agents themselves. Without such independent safeguards, supervisory agents could inadvertently introduce new errors, vulnerabilities or compliance challenges. See Note 4 on mitigating measures that will help ensure secure and constrained guardian agent operations. This layered approach, often called "defense-in-depth," is gaining traction to counter overreliance on any single oversight mechanism and ensure guardians themselves remain bounded and auditable.

Deep integration of guardian agents with identity and access management capabilities: Guardian technologies are increasingly integrating with existing enterprise agent identity and access management systems, but will also shift emphasis from static identity credentials to real-time behavioral monitoring. In the absence of a global agent registry, organizations will prioritize advanced agent profiling and anomaly detection capabilities, and will rely on metadata for fingerprinting in the absence of declared agent identities, to counter escalating threats such as privilege escalation, authorization bypass, and unauthorized data access. Guardian agents will use information in agent identity registries to enrich their own agent identifiers, which are often based on discovered metadata and fingerprints. Emerging frameworks treat AI agents as high-privilege, nonhuman identities requiring continuous behavioral monitoring and just-in-time access controls to address gaps in traditional IAM systems.

Integrated agent IAM and information governance: A key trend is the convergence of agent identity, credential, and access management (ICAM) with information governance. The traditional separation between identity and data governance is narrowing, as forward-looking organizations manage these as integrated capabilities. Guardian agents are central to this alignment, ensuring robust identity and data controls to counter emerging threats and potential data compromise. This convergence is driven by the need for data-centric policies that govern both human and agent access, using tools like data activity monitoring to detect risky behaviors tied to sensitive information. By 2027, over 70% of AI agent identity providers will classify the sensitivity of data agents interact with as part of granting and scoping access rights.

Guardian agent integration in multiagent systems: Guardian agents are foundational to the trust, risk and security architecture of cross-platform and multiagent systems. Their integration with zero-trust frameworks and broader security and governance platforms addresses vulnerabilities such as rogue AI caused by malicious actors or benign mistakes, and supply chain attacks, strengthening enterprise risk mitigation. Guardian agents need to understand the intentions of individual goals that agents have, along with the intended collective goals of multiple agents interacting in a network across multiple hosted environments. Actions and outcomes must be aligned with the AI agent network's collective goals and intentions.

Independent guardian agents disrupt legacy security by 2029: By 2029, independent guardian agents will eliminate the need for almost half of incumbent security systems intended to protect AI agent activities today in over 70% of organizations. This prediction reflects the rapid maturation of agentic AI ecosystems, where autonomous guardian agents themselves natively handle agent identity management, dynamic data classification and sensitivity mapping, real-time monitoring and cleanup of agent permissions plus clear visibility into what sensitive data or systems each agent could potentially access or impact, contextual risk assessment, and runtime adaptation — continuously fusing threat intelligence with internal/external signals into enriched feeds for detection and response.

Equipped with this self-contained intelligence, independent guardian agents deliver native protection that combines deterministic methods (explicit policy engines, rule-based guardrails, and verifiable controls) with AI-driven reasoning to monitor, constrain, and optimize every agent action at runtime without relying on external systems for data classification, identity management and other security functions. gateway layers. This will render many current security layers redundant.

Addressing integration and communication challenges: The proliferation of agents is driving demand for advanced gateways and orchestration tools. Guardian agents require secure, standardized multiagent communication, with integration maturity expected to advance significantly over the next 12 to 24 months. Enterprise surveys show rapid progress in orchestration tools for multiagent workflows, enabling secure delegation and coordination while reducing fragmentation in hybrid deployments.

Regulatory and policy drivers: Increasing regulatory mandates around transparency and algorithmic guardrails are making robust guardian agent deployment essential for compliance, particularly in regulated sectors. These requirements are shaping global technology policy and driving enterprise adoption of advanced governance solutions. For example, Wall Street financial institutions will adopt emerging guardian agents from providers that monitor employee conversations on conferencing platforms like Zoom, which spawn agents to take actions on behalf of humans. Laws such as California's frontier AI transparency requirements (effective 2026) and the EU AI Act's high-risk system obligations (phased in through 2026) mandate enhanced oversight, risk assessments, and incident reporting that align directly with guardian agent capabilities in sensitive domains.

Market Analysis

The guardian agent market — encompassing technologies for the oversight, security, and governance of autonomous AI agents — is entering a phase of accelerated growth, underpinned by the rapid adoption of agentic AI across industries. By 2030, guardian agent solutions are forecast to account for at least 6% of the agentic AI market, translating to an annual value of over $3 billion as the broader AI agent sector approaches $52.62 billion with a 46.3% CAGR from 2025. See Markets and Markets Market Reports.

This expansion is driven by *heightened enterprise concern over new risks introduced by autonomous agents, including aberrant behavior, adverse business outcomes, privilege escalation and data leakage*, which are prompting organizations to invest in robust oversight and governance layers.

> **AI agents simply can't be trusted to follow instructions as intended — making them unreliable and impossible to depend on. Use guardian agents to deliver essential trust, risk and security capabilities and to ward off adverse outcomes from aberrant behavior and new cyberthreats. And make sure you Guard the Guardians themselves.**
>
> — *Source: Gartner (February 2026)*

Key dynamics shaping the market in 2026 include:

- **Adoption acceleration:** Seventeen percent of surveyed CIOs indicated their enterprise had already deployed AI agents, and another 42% planned to deploy them within one year, according to the 2026 Gartner CIO and Technology Executive Survey conducted in June 2025. This will necessitate guardian agent solutions to mitigate risks, particularly in multiagent environments. [1]
- **Shift to proactive governance:** The market is evolving from reactive security models toward proactive governance, with integration into zero-trust frameworks and a focus on behavioral monitoring rather than static controls.

- **Disruption of organizational management structures:** AI agents demand clear organizational accountability and an AI leader, and new cross-enterprise team processes and tools for effective runtime deployments. This ensures all interested parties work together in achieving unified AI outcomes that meet their different requirements.

Gartner observes five organizational trends driving TRiSM and guardian agent oversight teams over the next three years, featuring hybrid structures including centralized AI leadership with decentralized units and dotted-line reporting to AI leaders:

- AI centralized or hybrid teams emerge under the leadership of an AI leader, and manage all aspects of AI, including technical controls.
- Business units have a strong vested interest in ensuring deployed AI agents meet business goals and align with specific objectives.
- Security and risk management functions coalesce and integrate with relevant IT software engineering and quality assurance teams, and shift left to AI engineering to manage technical controls.
- Fractures in information governance silos surface, causing a renewed immediate need for a unified approach across different organizations for protecting information used and generated by AI.
- Legal and compliance staff demand that regulatory and legal requirements are satisfied in enterprise AI applications.

Market fragmentation: The guardian agent market is highly fragmented, reflecting diverse solution focuses and buyer needs amid rapid AI advancements, with growing convergence between nimble startups and entrenched enterprise giants.

See Note 5 for emerging use cases for guardian agents.

Key Solution Provider Segments

The guardian agent market is segmented by the specific challenges and requirements of buyers, and also by what is most accessible and practical for them to adopt — whether that means choosing specialized new vendors or leveraging established enterprise platforms that are expanding into guardian agent capabilities. Further there is growing overlap, competition, and collaboration between agile startups and traditional providers.

- **Agent security and risk specialists:** Innovators in supply chain governance, posture management, threat detection, runtime protection, secure code generation, and/or customer facing digital channel protection, targeting advanced risks such as access abuse, prompt injections, agent hijacking, and tool exploits. These solutions appeal to organizations seeking cutting-edge, adaptable defenses.
- **Business alignment and outcome optimizers:** Solutions that align agent actions with business objectives, emphasizing outcome optimization and intent verification to ensure agents deliver value within preset strategic, tactical and compliance boundaries. They appeal to business owners of AI agent systems.
- **IT/Security platform vendors:** Established enterprise vendors are embedding Guardian capabilities into existing workflows, supporting reliable, large-scale protection for automation and analytics.
- **AI agent development and governance platforms:** Cloud-based platforms supporting the full life cycle of enterprise agents, with integrated guardian agent functions including monitoring, policy enforcement, and compliance tools.
- **Identity management vendors:** Providers specializing in agent identity and credential governance, applying zero-trust principles to manage agent access and prevent abuses such as authorization bypass. These vendors are starting to incorporate information governance capabilities for an integrated approach (see Note 9).
- **AI content governance:** Vendors specializing in governing AI-generated content and outputs for accuracy, compliance, brand consistency, ethical standards, and risk mitigation, including scanning, moderation, and rewriting tools.

Evaluation Criteria

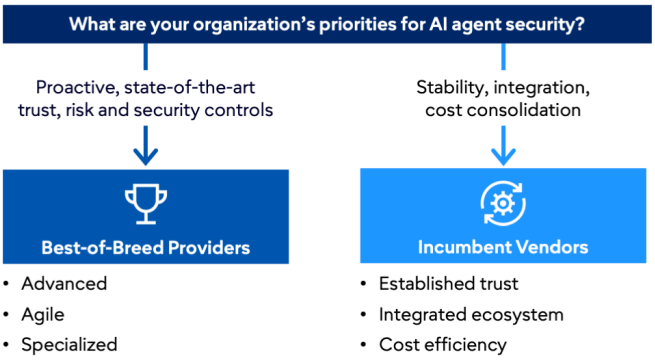
- Clients should consider picking a best-of-breed provider for guardian agent controls when they require proactive, agile, and specialized trust, risk and security measures to counter the sophisticated emerging threats and unpredictable behavior associated with AI agents. These providers offer advanced capabilities, such as real-time runtime inspection, dynamic policy enforcement, alignment of agent actions with user intentions, and automated exposure and posture management, making them particularly suited for organizations that must adapt quickly in a fast-evolving threat landscape.

- Conversely, incumbent vendors — a choice characterized by established trust and integrated ecosystems — may still be preferable for organizations that prioritize stability, extensive integration with existing IT systems, and cost consolidation. However, for organizations facing the most emerging and nuanced AI agent risks and threats, the inability of incumbent vendors to update as quickly with modern controls could leave critical gaps in protection. To address this, incumbent vendors often partner with agile startups to compensate for their lagging innovation, either through direct integration or by offering their services in their marketplace for third-party guardian agents.

Ultimately, the decision on which solution to use must be based on a detailed risk assessment, clear mapping of organizational needs, and a thorough review of a vendor's innovation roadmap and real-world performance evidence. As the guardian agent market consolidates and evolves, clients that make a proactive vendor selection today will be better positioned to derisk and secure their AI agents — and thus their overall enterprise — against tomorrow's risks.

Figure 4: Decision Tree for Guardian Agent Provider Selection

Decision Tree for Guardian Agent Provider Selection



Source: Gartner
836388

Gartner

See Table 2 in Note 6 for a decision framework for selecting a guardian agent provider.

Organizations adopting integrated, flexible guardian solutions — combining embedded and independent oversight — will be best positioned to secure competitive advantage and realize the full productivity potential of autonomous systems.

However, persistent challenges, including evolving threats and governance gaps, necessitate vigilant, adaptive strategies to unlock the potential value from agentic AI.

Representative Vendors

The vendors listed in this Market Guide do not imply an exhaustive list. This section is intended to provide more understanding of the market and its offerings.

Representative vendors are listed below per category, as summarized in the "market analysis" section. Vendors are diverse and fragmented in their guardian agent solution set. Prioritize your requirements and evaluate products for their ability to satisfy them by mapping your functional needs against the guardian agent capabilities listed in the Market Definition.

Note that this Market Guide excludes providers who do not provide native guardian agent controls in all three mandatory feature categories (see Market Definition). As such, it does not include vendors who partner with others to complement native offerings to complete their range of controls. Sample vendors here include OneTrust, Credo AI and Cranium, and frontier AI model providers and hyperscalers such as Anthropic, AWS and Azure OpenAI (see Note 7).

Vendor Selection

Table 1: Representative Vendors

(Enlarged table in Appendix)

Category	Description	Sample vendors
Risk and security specialists	Emerging companies specializing in dedicated AI/agent security, posture management, threat detection, and runtime defenses.	Aiceberg, Airrived, Alice, Apiiro, Capsule Security, CHEQ, Dtex, Galene.ai, Geordie, Holistic AI, Knostic, Lumia Security, NeuralTrust, Noma Security, Onyx Security, Opsin, Pillar, Portal26, Singulr AI, Straiker, Sun Security, Varonis, Vijil, Virtue AI, Xeris, Zenity
Business alignment and outcome optimizers	Startups focusing on business goals and aligning agent outcomes with the agent creator's and the organization's intent.	Avon.ai, ChatSee, Wayfound
Agent identity*	Vendors managing identities/access for AI agents with zero-trust policies and governance.	Astrix Security, BeyondTrust, Delinea, Entro Security, Microsoft Entra, Okta, Orchid Security, Palo Alto Networks (CyberArk), PlainID, Silverfort
IT or security platform vendors	Established enterprise workflow, IT service or security management platforms integrating guardian features for safe AI agent use in business processes.	Cato Networks (AIM), Crowdstrike, IBM (Watsonx governance), Palo Alto Networks (Protect AI), Sentinel One (Prompt Security), ServiceNow
AI agent development and governance platforms	Cloud/data platforms with governance layers for developing, deploying, and securely overseeing custom AI agents, featuring real-time monitoring, policy enforcement, content safety, observability, and compliance for safe autonomous agent scaling.	AgilePoint, Airia, AWS (Bedrock Guardrails), Databricks (Mosaic AI Gateway & Guardrails), Google Cloud (Vertex AI Agent Builder + guardrails & monitoring), Microsoft (Azure AI Content Safety + Agent 365), Salesforce (Agentforce)
AI content governance	Vendors specializing in governing AI-generated content for compliance, brand consistency, ethical standards, and risk mitigation, including scanning, moderation, and rewriting tools.	Bynder, Fujitsu LLM scanner, Markup.AI
*See Note 9		

Source: Gartner (February 2026)

Market Recommendations

The guardian agent market is fragmented and rapidly evolving, making it challenging for organizations to identify necessary capabilities and suitable vendors. No single provider, including major hyperscalers like Microsoft, Google, or AWS, offers a complete solution.

- Gartner recommends focusing on four key requirements and evaluating vendors against their ability to deliver on them: agent discovery, identity and access management, information governance, and policy enforcement. Visibility over all agents — sanctioned and unsanctioned — is the most critical starting point.

- Hyperscalers provide governance controls, but these are often limited to their own ecosystems and may not manage third-party agents effectively. Organizations should supplement hyperscaler controls with independent vendors and be mindful of potential overlap with existing IT and Security platform partners. Guardian agents are becoming essential, as traditional controls alone cannot address the growing risks from AI agent proliferation.
- When selecting vendors, organizations should conduct comprehensive risk assessments, evaluate innovation roadmaps, and prioritize evidence-based case studies. Integration with existing platforms and flexibility to adapt to market changes are important, especially given ongoing vendor consolidation trends.
- Finally, organizations must balance cost with innovative capability, considering both immediate expenses and long-term trust, risk and security needs. By following these recommendations, organizations can better navigate the complex AI agent security landscape and ensure robust protection.

Evidence

2026 Gartner CIO and Technology Executive Survey. This survey was conducted online from 1 May through 30 June 2025 to help CIOs and technology executives benchmark their priorities and investment plans against those of peers worldwide. Qualified respondents led a digital/technology function and were accountable for running or improving/growing a specific area of their enterprise. In total, 2,501 CIOs and technology executives participated, with representation from all geographies, revenue bands and industry sectors (public and private).

Note 1: Gartner's Initial Market Coverage

This Market Guide provides Gartner's initial coverage of the market and focuses on the market definition, rationale for the market and market dynamics.

Note 2: AI Agent Management Platforms (AMPs)

AI agent management platforms (AMPs) provide a unified interface to securely manage, monitor, govern, acquire, organize, and generate analytics on AI agents and their supporting toolsets. A key component of AMPs will be the ability to observe and guardrail AI agent activity. The integration of guardian agents to the AMP is an essential component of managing AI agents across an enterprise. See AI Vendor Race: AI Agent Management Platform: The Most Valuable Real Estate in AI for more details on AMP functions and architecture.

Note 3: Sample DIY Tools and Frameworks

Figure 5: Sample Do-it-Yourself Tools and Frameworks

Sample Do-it-Yourself Tools and Frameworks

Category	Purpose	Maturity and Ease of Use	Typical Integration Effort	Best For
AI Agent Platforms and Frameworks (CrewAI, AutoGen, LangGraph, Smolagents, SuperAGI...)	Build, orchestrate, and run autonomous or collaborative AI agents	Rapidly maturing; many open-source	Low to medium	Teams wanting to create multi-agent systems quickly; from prototypes to production
AI Observability (Dynatrace, Arize, WhyLabs, Langfuse, etc.)	Monitor LLM/agent performance, trace prompts/responses, detect drift & bias	Enterprise tools mature; open-source emerging	Medium to high	Organizations needing governance, auditability, and cost control of live agents
Data Observability (Monte Carlo, Bigeye, Soda, Anomalo, etc.)	Ensure quality & freshness of data produced or consumed by agents	Very mature in data engineering space	Medium	Data-centric teams treating agent outputs as first-class data products
DIY with Open Source (ELKI, Loglizer, Edge Delta, Grafana LGTM...)	Full-custom anomaly detection & observability using logs, metrics, APIs	Highly flexible, community-driven	High	Teams with strong engineering bandwidth who need total control and no vendor lock-in

Source: Gartner
836388

Gartner

Note 4: Metagovernance for Guardian Agents

Drawing on leading AI governance practices, Gartner identifies five critical controls to ensure guardian agents remain secure, reliable, and aligned with organizational objectives:

- **Contextual access control:** Treat guardian agents as unique service identities within IAM systems, assigning granular, context-aware roles and dynamically adjusting permissions based on data sensitivity and operational context. This enforces least privilege and prevents unauthorized access or overreach.
- **Input and output filtering:** Deploy input sanitization and output filtering to protect against prompt injection attacks and ensure compliance with content policies. This minimizes the risk of manipulation, data leakage, or biased oversight.
- **Task execution control and sandboxing:** Restrict guardian agent operations to sandboxed environments with whitelisted APIs, rate limits, dry-run simulations, and rollback capabilities. These controls prevent supervisory agents from disrupting critical processes or exceeding their authority.

- **Continuous observability:** Implement real-time monitoring of key metrics — such as intervention frequency and behavioral anomalies — and integrate alerts for rapid response. This ensures timely detection and remediation of any deviations in agent behavior.
- **Logging, traceability, and auditability:** Maintain immutable, timestamped logs of all guardian agent actions and decisions for audit and compliance purposes. This provides full traceability and accountability for supervisory activities.

By adopting these controls, organizations can maximize the benefits of guardian agents while ensuring their oversight systems remain trustworthy, transparent, and fully aligned with enterprise governance requirements.

Note 5: Emerging Use Cases for Guardian Agents

Guardian agents oversee semi- or fully autonomous AI systems for compliance, safety, and reliability. Sample emerging use cases are listed below.

For each of these use cases, domain-specific guardian agents will be deployed, with more granular sub-domain agents operating under the umbrella of each primary domain. These agents will be packaged into highly performant supervisory capabilities, designed to efficiently and accurately oversee other agents with low latency, leveraging tailored technical controls and evaluations to ensure effective, real-time governance.

- **Personal assistant assurance:** Guardian agents oversee multipurpose productivity AI agents handling tasks like research, scheduling, business-building and operations, and personal assistance, monitoring inputs/outputs in real time to ensure agent activities align with agent goals, and to prevent data leaks, biases, or inaccuracies while ensuring compliance and ethical standards. This promotes user trust by intervening on risks such as unauthorized data sharing or hallucinated information.
- **Developer ecosystem and code generation oversight:** Oversee AI agent developer environments including the IDE, Skills utilized, MCP server posture, interactions between the IDE agent and installed MCP servers. Oversee AI agents generating code, seeking to ensure code generates secure code, includes threat prevention, real-time detection and data inspection, blocking and remediation of vulnerabilities, bugs, and adherence to security and best practices.
- **Industry-specific applications:** Oversee AI agents in specialized operations across sectors, such as CRM systems, supply chain and logistics, healthcare diagnostics, and financial services.

- **Content review and governance:** Oversee AI agents in content and output creation and management processes, scanning, moderating, and enforcing standards for AI-generated outputs to ensure accuracy, brand consistency, ethical guidelines, regulatory compliance, and risk mitigation against issues like misinformation, hallucinations, inaccuracies or bias.

- **Communications governance:** Oversee AI agents in digital communication and collaboration platforms (e.g., using tools that integrate with UCC platforms such as Zoom), providing posture management, real-time risk detection, blocking and remediation for AI agent actions to prevent violations, ensure record-keeping, and flag anomalies especially in regulated environments.

Note 6: Decision Framework for Selecting a Provider

Table 2: Decision Framework for selecting a Provider

(Enlarged table in Appendix)

Factor	Best-of-breed niche provider	Incumbent vendor	Explanation
Innovation and feature set	Offers advanced, specialized features such as risk scored agent catalogues, alignment of agent intentions and dynamic runtime controls.	Provides traditional, stable features with occasional updates.	Niche providers rapidly incorporate emerging risks and threats and proactive measures while incumbents may lag in innovation. (See Govern AI Using TRiSM: The Technical Framework for Trust, Risk, and Security; AI Vendor Race: A Simple Foundation for AI TRiSM Product Success.)
Integration and ease of adoption	Designed for easy integration with modern AI architectures and offers robust customer enablement services.	Integrated with legacy systems and offers broad, established security ecosystems.	Clients preferring rapid deployment and specialized support may lean towards niche providers. (See Sensible AI Requires Service Providers to Ground Messaging in Demonstrable Value.)
Risk management and proactive capabilities	Proactively mitigates emerging AI-specific risks with automated exposure management and continuous oversight.	Delivers stable risk controls but may lack agility in response to new AI threats.	Proactive controls like real-time agent alignment enforcement, restricting agency of agents, and threat detection are critical for fast-evolving environments. (See Emerging Tech: AI Vendor Race: Future-Proof MDR With Automated Exposure Management and Zero Trust.)
Economic considerations	May command a premium but usually deliver faster ROI with specialized solutions.	Often bundled with other services, offering cost efficiencies but with potential risk of lock-in.	Evaluating total cost of ownership is essential, especially when rapid innovation is needed. (See Sensible AI Requires Service Providers to Ground Messaging in Demonstrable Value.)

Note 7: Hyperscalers and AI Agent Development and Governance Platforms

AI agent development and governance platforms represent a broad category of vendors ranging from the hyperscalers (Microsoft, Google, AWS) to AI data platforms (Snowflake, Databricks, etc.), to AI orchestration startups (such as Airia) that offer agent building, orchestration, management and governance features.

While these solutions are diverse in terms of their origins and use cases (e.g., M365 and Google as productivity platforms, Snowflake and Databricks as big data analysis solutions), what they have in common is a focus on providing a secure boundary for AI agent activity. The strategy here is to become the vendor of choice for managing what will likely be the most important new piece of IT infrastructure in a generation — the AI agent.

Microsoft is pushing hard in this space with the arrival of Agent 365 (currently in Preview). The promise is that as long as you register your agent, (native or third party) with Microsoft Entra ID, then you will benefit from the security and governance controls that come with the wider Microsoft platform (Purview, Entra, Defender, etc.).

However, there is a lock-in risk here as the vendor will not only provide the model access and building tools for these agents, but will also provide the management layer as well. Gartner has yet to see evidence that hyperscaler AI governance solutions can fully manage third-party agents created outside of the environment in a comparable way to first party agents. It is also unlikely that rival AI providers will want to outsource management of their AI agents to their competition, leading to gaps in coverage provided.

As such, Gartner believes that organizations must ensure their guardian agents work equally well across platforms and complement existing agent development and governance platforms with independent AI security and governance providers.

Note 8: Evaluation Methods Used by Guardian Agents

To yield the most accurate results, guardian agents should start with the most efficient deterministic evaluation (and typically lowest cost) method available when assessing agent interaction risk, before turning to an LLM or SLM for judgment. In practice, this means guardian agent runtime evaluations start by:

- Using preestablished rules,
- Then progress to behavior monitoring using statistical analysis and contextual evaluation

- And only then move to an LLM and SLM if initial evaluations using the preceding methods are inconclusive.

There are situations, however, where the GA evaluation will jump straightaway to LLM or SLM judgment, for example under these scenarios (for more on this topic, see Guardians of the Future: How CIOs Can Leverage Guardian Agents for Trustworthy and Secure AI):

- **Complex context:** When the intention involves nuance or ambiguity (e.g., "create a phishing email" for a training demo vs. malicious activity), LLM evaluation is needed to interpret context beyond simple rules.
- **Risk indicators:** Upfront metadata — like a user's history of flagged behavior or a high-risk transaction (e.g., "transfer funds to unverified account") — triggers LLM scrutiny if deterministic steps can't connect the dots.
- **Urgency and impact:** High-stakes cases (e.g., "execute code on production server") with potential for harm or disruption bypass deterministic checks for immediate LLM analysis.
- **Insufficient deterministic capabilities**: If basic filters (e.g., keyword matching) can't judge the full scope — like intent, scale, or downstream effects — the guardian agent escalates to LLM.
- **Efficiency trade-off**: When prior intel (e.g., system alerts) suggests deeper scrutiny is inevitable, skipping to an LLM saves time over running redundant simple checks.

Also, refer to OWASP Agent Observability Standard for evolving industry standard on observability.

Note 9: Identity Management and Information Governance

An emerging trend is that agent IAM and information governance are becoming closely connected and should be overseen as integrated functions. Integrating agent IAM with information governance allows organizations to establish unified oversight of both identity and data usage. This approach ensures that AI agents are not only authenticated and authorized properly but also that their activities are auditable and compliant with regulatory requirements such as NIST standards and zero trust frameworks.

Forward-thinking organizations are recognizing that a siloed approach to identity and data governance is no longer sufficient. By aligning these functions, they can streamline compliance processes, improve operational efficiency, and provide a more predictable and secure environment for both human and AI-driven interactions. This integrated strategy is increasingly seen as essential for managing the complex trust relationships and data flows that characterize multiagent ecosystems.

As organizations move toward integrating identity and information governance, the need for robust solutions that facilitate this convergence becomes increasingly important.

Table 1 (Vendor table) highlights sample technology providers in the Identity category who are gradually incorporating stronger information governance capabilities to support an integrated approach and agent security assurances. This is often accomplished through partnerships they establish with information governance providers that focus on data classification, sensitive data discovery, access context, permissions hygiene, and preventing sensitive data exposure to AI agents/copilots/LLMs. Information governance solutions can likewise be integrated with other categories of GA providers in this table.

These information governance features are important “add-ons” to protect enterprises from risks like oversharing sensitive data with AI systems. Sample vendors in the information governance category that complement agent identity and other GA solutions include Cyera, Bigeye, Concentric AI, Touchdown, and Collibra.

Further, while these vendors originated — and largely continue to operate primarily — as information governance or data-security platforms, most are aggressively expanding into dedicated AI and agent-specific capabilities in direct response to the rapid surge in AI agent adoption. Their expansion includes evolving capabilities in agent discovery and inventory, contextual risk mapping (linking each agent to the exact sensitive data, identities, permissions, and information it can access), and real-time runtime enforcement.

Recommended by the Authors

Some documents may not be available as part of your current Gartner subscription.

AI Vendor Race: AI Agent Management Platform: The Most Valuable Real Estate in AI

Act Now: Take These 5 Steps for AI Agent Assurance

The Current State of AI Agents for Enterprises

Guardians of the Future: How CIOs Can Leverage Guardian Agents for Trustworthy and Secure AI

Market Guide for AI Trust, Risk and Security Management

Market Guide for AI Governance Platforms

Moltbots Need Guardian Agents: What Are They?

Table 1: Representative Vendors

Category	Description	Sample vendors
Risk and security specialists	Emerging companies specializing in dedicated AI/agent security, posture management, threat detection, and runtime defenses.	Aiceberg, Airrived, Alice, Apiiro, Capsule Security, CHEQ, Dtex, Galene.ai, Geordie, Holistic AI, Knostic, Lumia Security, NeuralTrust, Noma Security, Onyx Security, Opsin, Pillar, Portal26, Singulr AI, Straiker, Sun Security, Varonis, Vijil, Virtue AI, Xeris, Zenity
Business alignment and outcome optimizers	Startups focusing on business goals and aligning agent outcomes with the agent creator's and the organization's intent.	Avon.ai, ChatSee, Wayfound
Agent identity*	Vendors managing identities/access for AI agents with zero-trust policies and governance.	Astrix Security, BeyondTrust, Delinea, Entro Security, Microsoft Entra, Okta, Orchid Security, Palo Alto Networks (CyberArk), PlainID, Silverfort
IT or security platform vendors	Established enterprise workflow, IT service or security management platforms integrating guardian features for safe AI agent use in business processes.	Cato Networks (AIM), Crowdstrike, IBM (Watsonx governance), Palo Alto Networks (Protect AI), Sentinel One (Prompt Security), ServiceNow
AI agent development and governance platforms	Cloud/data platforms with governance layers for developing, deploying, and securely overseeing custom AI agents, featuring real-time monitoring, policy enforcement, content safety, observability,	AgilePoint, Airia, AWS (Bedrock Guardrails), Databricks (Mosaic AI Gateway & Guardrails), Google Cloud (Vertex AI Agent Builder + guardrails & monitoring), Microsoft (Azure AI Content Safety + Agent 365), Salesforce (Agentforce)

	and compliance for safe autonomous agent scaling.	
AI content governance	Vendors specializing in governing AI-generated content for compliance, brand consistency, ethical standards, and risk mitigation, including scanning, moderation, and rewriting tools.	Bynder, Fujitsu LLM scanner, Markup.AI
*See Note 9		

Source: Gartner (February 2026)

Table 2: Decision Framework for selecting a Provider

Factor	Best-of-breed niche provider	Incumbent vendor	Explanation
Innovation and feature set	Offers advanced, specialized features such as risk scored agent catalogues, alignment of agent intentions and dynamic runtime controls.	Provides traditional, stable features with occasional updates.	Niche providers rapidly incorporate emerging risks and threats and proactive measures while incumbents may lag in innovation. (See Govern AI Using TRiSM: The Technical Framework for Trust, Risk, and Security; AI Vendor Race: A Simple Foundation for AI TRiSM Product Success.)
Integration and ease of adoption	Designed for easy integration with modern AI architectures and offers robust customer enablement services.	Integrated with legacy systems and offers broad, established security ecosystems.	Clients preferring rapid deployment and specialized support may lean towards niche providers. (See Sensible AI Requires Service Providers to Ground Messaging in Demonstrable Value.)
Risk management and proactive capabilities	Proactively mitigates emerging AI-specific risks with automated exposure management and continuous oversight.	Delivers stable risk controls but may lack agility in response to new AI threats.	Proactive controls like real-time agent alignment enforcement, restricting agency of agents, and threat detection are critical for fast-evolving environments. (See Emerging Tech: AI Vendor Race: Future-Proof MDR With Automated

			Exposure Management and Zero Trust.)
Economic considerations	May command a premium but usually deliver faster ROI with specialized solutions.	Often bundled with other services, offering cost efficiencies but with potential risk of lock-in.	Evaluating total cost of ownership is essential, especially when rapid innovation is needed. (See Sensible AI Requires Service Providers to Ground Messaging in Demonstrable Value.)
Enterprise scale and ecosystem	Best for organizations that need agility, rapid innovation, and specialized security controls.	Suitable for enterprises relying on existing broad security ecosystems with known vendors.	Large organizations may prefer incumbents for standardized integration and steady support; however, the rapidly evolving needs of guardian agents favor niche solutions. (See Market Guide for AI Trust, Risk and Security Management.)

Source: Gartner (February 2026)