Geordie

MARKET ASSESSMENT

What the agent security market watches, and what it still can't see

A market assessment for security leaders evaluating agentic AI risk approaches.

A guide for security leaders evaluating agentic AI risk.

Published September 2026.

Table of contents

IN BRIEF

Model vendors describe the agent as a configuration of their own platform. Identity vendors describe it as a credential to verify. Gateway vendors describe it as traffic to inspect. Each is instrumenting the layer it already operates in, not the layer where agentic risk actually accumulates.

That layer is behavior: the sequence of actions an agent takes once it's running. Most organizations can't yet tell human activity from agent activity in their own logs, and fewer than half of deployed agents are monitored at all. Closing the gap means treating the agent as the unit of analysis, and behavior, not identity or traffic alone, as what gets instrumented.

What the evidence shows

- Identity dominated the response: RSA Conference 2026's wave of agent-security launches centered on verifying who an agent is, not what it does.
- Each cluster extends the surface it already watches, since that's what its platform already logs or sees: the model for model vendors, credentials for identity vendors, the boundary for gateway vendors.
- The instrumented layer is a single point in time. A credential check and a gateway inspection are both, structurally, one moment. Agent risk is a sequence that accumulates over a session.
- The gap is measured now, not theoretical. Independent research from the Cloud Security Alliance and Gravitee both finds that most organizations cannot see, and are not securing, the majority of their own agent activity.

THE NUMBERS THE MARKET IS PLANNING AROUND

33%

Of enterprise software applications to include agentic AI by 2028 (Gartner)

15%

Of day-to-day operations forecast to be completed autonomously by 2028 (Gartner)

40%

Of agentic AI projects forecast to face cancellation over unmanaged risk (Gartner)

On method. Figures are drawn from third-party research and public analyst forecasts, attributed at first use. Nothing here is a Geordie customer figure.

01 · How the market has organized itself

Every AI platform vendor now ships an agent. Every security vendor now ships an agent control. The two groups are not securing the same layer of the same problem.

Model-platform vendors treat the agent as a configuration state of the model: skills, connectors, and tool access are extensions of its capability. Governance appears at the ecosystem's edge, as partner assurance, not a property of the agent's own runtime behavior.

Security vendors go further, rightly: an agent has identity, memory, and a sequence of actions the model alone doesn't capture. But most of the market's response has concentrated on a single layer: identity.

At RSA Conference 2026, the wave of agent-security launches centered on identity frameworks: verifying who an agent is, where it came from, what credentials it carries. Post-conference analysis converged on one finding: these frameworks verified identity; few tracked behavior.

A second cluster inspects traffic at the network and API boundary, the way a firewall inspects packets, reusing infrastructure security teams already operate. It's a defensible start, but by construction external to the agent: it sees what crosses the boundary, not what happens inside the decision.

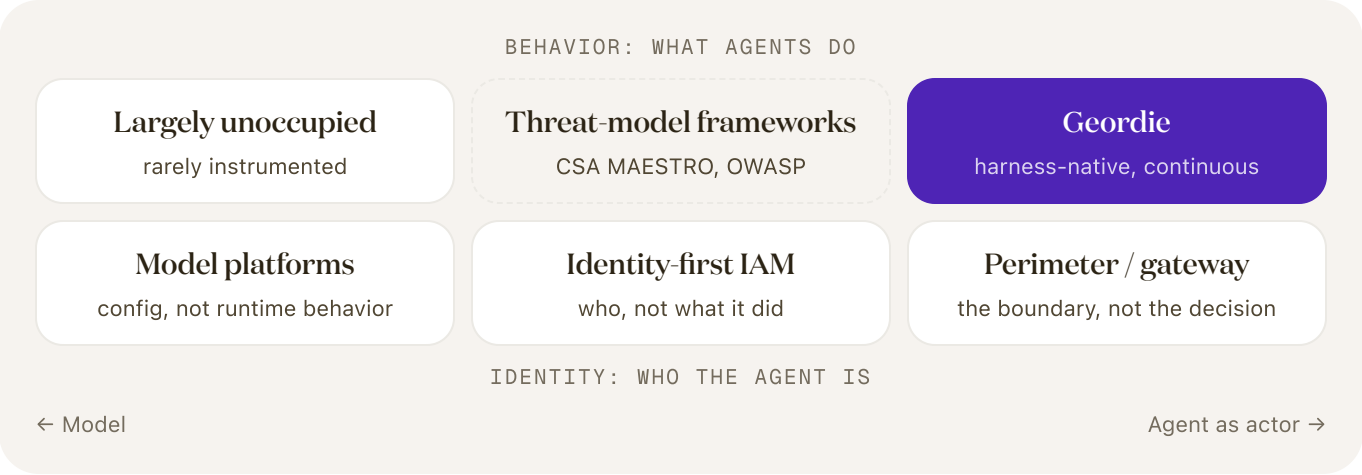

Figure 1. Six clusters, mapped by what each treats as the unit of control and what it instruments about it. Dashed outlines mark standards and internal practice, not vendors; solid purple marks a vendor product.

The pattern is not random. Each cluster extends the surface it was already built to observe. Identity vendors extend identity, because credentials are what their platforms already log. Gateway vendors extend inspection at the boundary, because traffic is what their platforms already see. Model vendors extend the model, because the model is the product. None of this reflects poorly on any one vendor's engineering. It reflects how markets form around existing instrumentation, rather than around the shape of the new problem.

Not everything on this map is a vendor, and the distinction matters when evaluating what to adopt. Threat-modeling frameworks such as CSA MAESTRO and OWASP's agentic guidance are published, vendor-agnostic standards. Nobody sells them, and no single company controls their roadmap. Manual governance isn't a standard either. It's what an organization does by default before it adopts a standard or a product: a spreadsheet inventory, a periodic policy review. Model-platform controls, identity-first IAM, perimeter and gateway inspection, and Geordie are the four vendor product categories on this map, the ones actually evaluated, licensed, and deployed. A framework can inform what's required of a product. It isn't a substitute for one, and it was never trying to be.

The top-left quadrant, model-centric and behavior-focused, stays empty for a structural reason more than an access reason. Behavior, in the sense that matters for risk, is a sequence of actions taken in an environment: calling a tool, reading a file, sending a request. A model on its own doesn't do any of that. It produces an output. Nothing worth calling behavior happens until a model is embedded in a harness with tools and a loop, and at that point what's behaving is the agent, not the model.

Closed model weights compound the gap. They are not the root cause of it. Weight access buys interpretability: some ability to reason about why a model produced a given output. It does not buy behavioral observability: a record of what a deployed instance actually did across a session. An open-weight model, dropped into the same harness, would leave the quadrant just as empty, because the gap was never about what's inside the model. It was about the fact that nothing worth calling behavior exists before the model starts acting.

What little occupies this space is static rather than continuous: model cards and safety evaluations published at release, characterizing how a model behaves in aggregate against a fixed benchmark. That's a snapshot of tendency, not a record of what happened. It answers a different question than runtime observability does, which is why it doesn't close the gap.

02 · Where each framing runs out

Identity answers a narrow question well: is this the agent it claims to be. It does not answer what the agent did once authenticated, in what order, or why. Posture tells you what an agent is configured to do. Behavior tells you what it actually does.

Figure 2. A credential check and a gateway inspection are both, structurally, a single point in time. Behavior is a sequence.

There is a second cost that shows up later: drift. Tools get added, prompts get modified, permissions get widened, usually by someone other than whoever ran the original review. A framework or inventory built at design time has nothing to say about a system that has since changed underneath it.

The gap is measured now, not theoretical. A Cloud Security Alliance and Aembit survey found 68% of organizations can't distinguish human from AI agent activity in their own logs; a separate 2026 CSA study of 235 security leaders found 92% lack full visibility into their AI identities, and 86% don't enforce access policies for them at all.

Figure 3. Selected findings from third-party research on agent security programs, 2026. Sources: Cloud Security Alliance & Aembit; Cloud Security Alliance (235 enterprise security leaders); Gravitee, State of AI Agent Security.

Perimeter and gateway inspection has a related but distinct limit. It is well suited to blocking a known-bad request at a chokepoint. It has less to say about a sequence of individually reasonable actions that accumulate into an unreasonable outcome: an agent that requests ordinary tool access repeatedly, in a pattern no single request would flag. Gravitee's 2026 State of AI Agent Security report puts the scale of the resulting blind spot plainly. Only 47.1% of deployed AI agents are actively monitored or secured at all.

None of this makes identity or perimeter controls wrong. It makes them necessarily incomplete once the unit of risk is a sequence of actions over time, rather than a credential check or a network packet.

03 · What each approach sees, and what it misses

Set side by side, the dominant approaches to agent governance describe a market that has instrumented the edges of the problem and left the middle largely dark.

Category	Type	Sees	Misses
Model-platform controls	Vendor product	Configuration and tool access granted to the model	Runtime behavior once tools are in use, and risk that spans more than one platform
Identity-first access management	Vendor product	Who the agent is: credentials, provenance, authentication	What the agent did after authenticating; sequence and drift over time
Perimeter and gateway inspection	Vendor product	Individual requests crossing a network or API boundary	Risk that accumulates across a sequence of individually reasonable actions
Threat-modeling frameworks	Vendor-agnostic standard	A structured taxonomy of where agentic risk can occur, e.g. CSA MAESTRO, OWASP	Runtime telemetry. Frameworks describe risk categories; they don't observe a live system
Manual governance	Internal practice	A point-in-time inventory or policy review	Continuous change as tools, prompts, and permissions are modified in production

Figure 4. What each governance approach sees and where it stops, mapped across five categories.

Read down the Misses column: each category is precise about a narrow slice and silent about the rest. The two non-vendor rows miss differently, since a framework or manual review is a point-in-time artifact by design. Neither gap is a flaw in any product; it's what happens when governance follows an existing boundary rather than the complete system an agent is.

04 · Timing and architecture

Why the timing matters

Gartner projects that by 2028, 33% of enterprise software applications will include agentic AI, with 15% of day-to-day operations completed autonomously. Gartner also predicts that 40% of agentic AI projects will face cancellation due to unmanaged risks.

None of this requires alarm. It requires sequencing. Agent estates are still forming inside most organizations, which means the instrumentation put in place now will still be describing the estate two years from now, not preserving a snapshot of how it looked on day one. The window for getting that instrumentation right is while the infrastructure is still forming, not after it has hardened around a partial view.

Architecture matters as much as coverage

Coverage is necessary. It isn't sufficient on its own, because how a control is enforced changes what it can and can't do.

Interception and enforcement are often treated as one mechanism: the gateway that inspects traffic is also the thing that blocks it. That coupling is a reasonable engineering choice. It also has a structural cost. A control sitting at an external boundary can only act on what crosses that boundary, in whatever form the boundary happens to observe: a network request, an API call, a tool invocation. It has no native view of the agent's own reasoning, memory, or the tool calls that never leave the agent's own runtime.

A harness-native control works differently. It sits inside the agent's own execution, applied through the agent's own configuration and context rather than a separate chokepoint. Enforcement happens at the point where the agent is actually deciding, not after the decision has already been made.

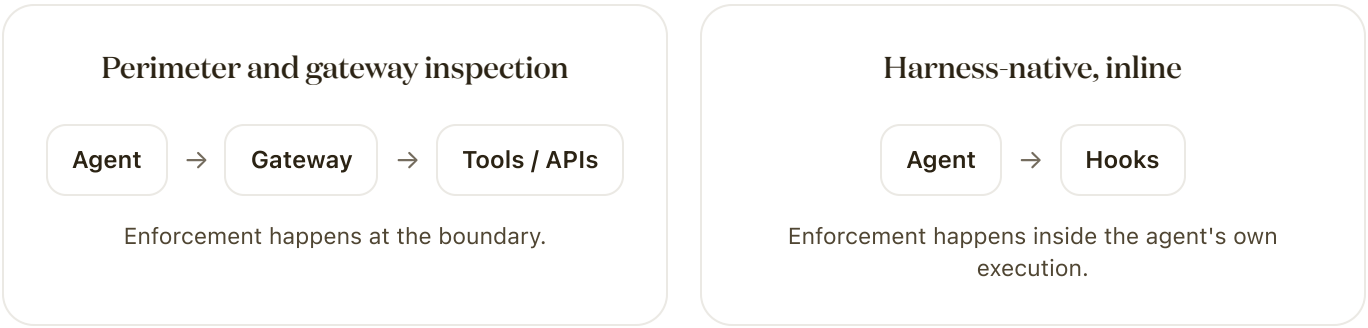

Figure 5. Interception and enforcement are separate concerns. Where a control sits changes what it can see and what it can do.

None of this makes gateways wrong. Removing them isn't the argument being made here. Agents are systems, not surfaces. A complete view of a system usually needs controls at more than one layer. Geordie works alongside existing gateways and platform-native controls. The architecture described here is additive. It is not a replacement for the boundary layer most security teams already operate.

05 · What this means for evaluation

Five questions separate a tool that reports on agents from a tool that reports on what agents do.

1. **What unit is actually being measured.** A tool that reports on identity or configuration is not reporting on behavior, even when its dashboard uses that word.
2. **Whether risk is inferred from a single action or a sequence.** A single tool call is rarely the problem. The accumulation of ordinary-looking actions over a session usually is.
3. **Where enforcement actually happens.** A control applied at an external boundary and a control applied within the agent's own context are different architectures with different failure modes. They are not interchangeable descriptions of the same thing.
4. **What happens to agents the inventory doesn't know about yet.** Agent estates are discovered continuously, not counted once. A static inventory is out of date on the day it is produced.
5. **Whether the answer changes as the agents change.** Tools, prompts, and permissions are modified in production constantly. Governance built for a point-in-time review will not hold across that drift.

This is the layer Geordie's platform is built to instrument: behavior, observed directly from the agent's own operating context, across cloud, code, and endpoint. It works alongside the identity and perimeter controls security teams already run.

Closing observation

Autonomy is arriving whether or not the instrumentation catches up to it. The gap this paper describes won't close on its own, and every quarter it stays open is a quarter of agent activity nobody can fully account for. The organizations who start instrumenting behavior now, while their agent estates are still forming, will be the ones who can answer for what their agents did, not just what they were configured to do.

SOURCES

RSA CONFERENCE 2026	Post-conference analysis of agent-security product launches
GARTNER	Forecasts for enterprise agentic AI adoption, autonomy, and project cancellation through 2028
CSA & AEMBIT	Survey on distinguishing human from AI agent activity in security logs
CSA, 2026	Study of 235 enterprise security leaders on AI identity visibility and access policy enforcement
GRAVITEE	2026 State of AI Agent Security report
CSA MAESTRO & OWASP	Vendor-agnostic agentic threat-modeling guidance

Geordie

geordie.ai