Mapping to the seven MAESTRO layers

SentinelleIA against the Cloud Security Alliance threat model for agentic AI systems, published July 2025.

Why this document exists

Agentic AI security is crowded with point tools, each covering one risk. MAESTRO is useful precisely because it refuses that framing: it decomposes an agentic system into seven layers, from the model itself up to the ecosystem of agents that call each other. A vendor that covers one layer is a feature. The table below shows where SentinelleIA sits across all seven.

Nine agents plus a supervisor. The supervisor is deliberately named apart, because the governance of the guardian is itself a requirement, not an implementation detail. It arbitrates authorizations and keeps its own audit chain, separate from the agents it governs.

The mapping

Layer	MAESTRO domain	What the layer covers	SentinelleIA coverage
L1	Foundation Models	Model level risks: jailbreaks, prompt level manipulation, unsafe or manipulated outputs.	Prompt Guard Agent, LLM Security Agent
L2	Data Operations	Data pipelines, retrieval sources, third party datasets and model dependencies.	Supply Chain Agent, Tool Protection Agent
L3	Agent Frameworks	The agent runtime itself: reasoning loop, tool invocation, mandate enforcement.	Supervisor (authorization boundary), AI Firewall Agent
L4	Deployment Infrastructure	Where agents run and how they reach the outside world.	Gateway Agent, Tool Protection Agent
L5	Evaluation and Observability	Measurement, telemetry, detection of drift and of blind spots.	Visibility Agent, metrics and alerting chain
L6	Security and Compliance	Controls, evidence production, regulatory alignment.	Governance Agent, SHA-256 chained audit log
L7	Agent Ecosystem	Multi agent interaction, delegation, cascade failure between agents.	Resilience Agent, supervisor

Two components are cross cutting and therefore appear on more than one layer: tool protection on L2 and L4, and the supervisor on L3, L6 and L7. Every one of the nine agents is placed, none is orphaned.

What this document does not claim

Covering a layer means having a control that operates there at runtime, not exhausting every risk the layer contains. Tool validation, scopes, audit trails and non human identity are not new in themselves and other vendors ship them. What is specific to SentinelleIA is that four checks, tool and agent validation, action budget, logged attestation and trust level, form a single indivisible transaction rather than four sequential checks, and that all of it runs on your infrastructure with no outbound dependency.

On evidence production, the rules, the chained audit log, the compliance reports and the model cards exist and run today. Industrialising this into end to end automatic evidence generation is the object of our TRL 6 to 8 maturation. The trust matrix already carries the Article 14 state for human oversight; the oversight server itself is on the roadmap for the second half of 2026.

From framework to measurement

A mapping is a claim. The measurement is what makes it verifiable. On a bench derived from AgentDojo, we measure the attack success rate of a system in direct mode, then under SentinelleIA. The reference series of 22 August 2026, three runs on the gateway build in service, gives 0.85 direct against 0.00 shielded: a median with a range of 0.75 to 0.95 across the three runs, 17 of 20 attack combinations landing without protection and none with it. No legitimate task was blocked, and utility was identical in both modes. That last figure is published on purpose: blocking everything is

easy and useless, so the cost in legitimate work refused belongs next to the rest, including when it is not zero. Twenty combinations is a small suite, so this is a result, not a guarantee, and the figures move with the build they were measured on.

The first level of our assessment is free, requires no access to your systems, and returns an exposure map in about five days. sentinelleia.fr/evaluation