

Determinism for AI agents.

AI agents are moving from systems that generate information to systems that take actions. Soterias makes an agent's authority deterministic while the agent itself stays probabilistic, so it can act autonomously without the ability to take an unauthorized action.

The Problem. Frontier models are improving faster than the controls around them. Agents now write and execute code, query databases, call APIs, and delegate work to other agents. A bad model output can become an external action with consequences. Current approaches lean on supervising agents or a human in the loop, which defeats the point of autonomy and leaves defenders chasing a widening attack surface.

The Gap. Most agent security today works by influencing or observing model behavior: model guardrails, sandboxing, monitoring, human approval, or supervising agents. These remain useful, but none of them answers the fundamental question. What is the system technically capable of doing when the model behaves incorrectly?

The Thesis

Soterias builds deterministic guardrails for agents that allow full autonomy while preventing unauthorized actions at the enforcement layer. The agent may reason probabilistically about how to complete a task. Its permissions, delegation, and every security-critical action are evaluated deterministically, outside the model, before anything runs. Because the boundary is code, the guarantee is a certainty rather than a probability: an agent inside it cannot take an unauthorized action.

The agent proposes. The control plane decides.

Why Now?

The OpenAI and Hugging Face incident set a clear benchmark for where agentic security is heading. Frontier models will keep advancing faster than the mechanisms meant to contain them, and sandboxes, supervision, and similar controls will not be enough to stop unintended actions or attacks. For regulated industries to trust an agent, its behavior has to be provable, and provability is at odds with the probabilistic nature of a language model. Deterministic guardrails resolve that tension: the agent acts on its own while its behavior stays inside defined security and compliance policies, in any environment.

Case Study: Fintech

Adoption of language models in fintech is high compared with other industries, but our conversations show it is heavily segmented: most usage sits in assistive chatbots, not in agentic workflows built to automate. Agents are still not trusted with sensitive data because of security and compliance risk. Soterias provides the full stack a team needs to deploy multi-agent systems safely, from authorization and policy enforcement to execution control.

Determinism Benchmarking

Alongside enforcement, Soterias provides a determinism benchmark so a team can show it is fit to operate in the most regulated industries. The benchmark measures whether the security-critical constraints on an agentic workflow hold consistently across different agent behaviors. With it comes the tooling to raise a score: mechanisms to strengthen controls as frontier models develop, guidance on mitigating future capability, and the components that make the security-critical parts of a workflow deterministically constrained. Together, this is a complete service for making a workflow deterministic where it matters.