AGNTCon + MCPCon Japan

Designing Trust Boundaries for Agent-to-Agent Systems

Lessons learned from a technical PoC

Ryuji Iijima
Next Gen Service Platform Division, SoftBank Corp.

SoftBank

2026.09.10 Total 21 pages

About Me

Ryuji Iijima
飯島 竜児

SoftBank Corp.
Engineer

2020 — Joined SoftBank
DevOps Infrastructure Development

2021 - 2022 — Authentication Systems for Mobile Communications
- SoftBank Wi-Fi Spot
- ISP
- 4G/5G

2023 — Cloud PBX and FMC
- ConnecTalk

2024 — R&D of DataSpace

2025 — **Governance Platform for AI Agents**

About Our Products — AgentSecOps PF (tentative)

AI Agents Orchestrated Across Platforms to
Make AI Visible, Traceable, and Controllable

Orchestrate distributed AI agents across your organization to ensure AI is used safely and sustainably.

Make AI Visible
Visualization

- AI Agents
- Owners
- Usage Status

Make AI Traceable
Traceability

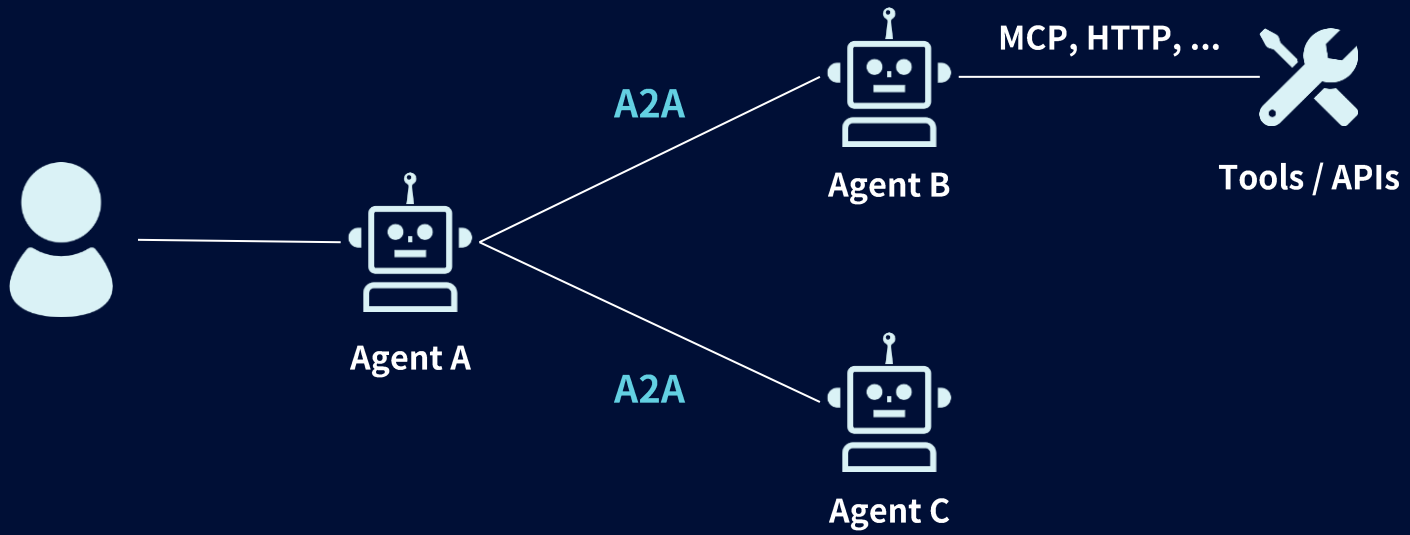
- Execution History
- Audit Logs
- Policy Decision Rationale

Make AI Controllable
Control

- Risky Executions
- Unauthorized Access
- Policy Violations

https://www.softbank.jp/sbnews/entry/20260422_01

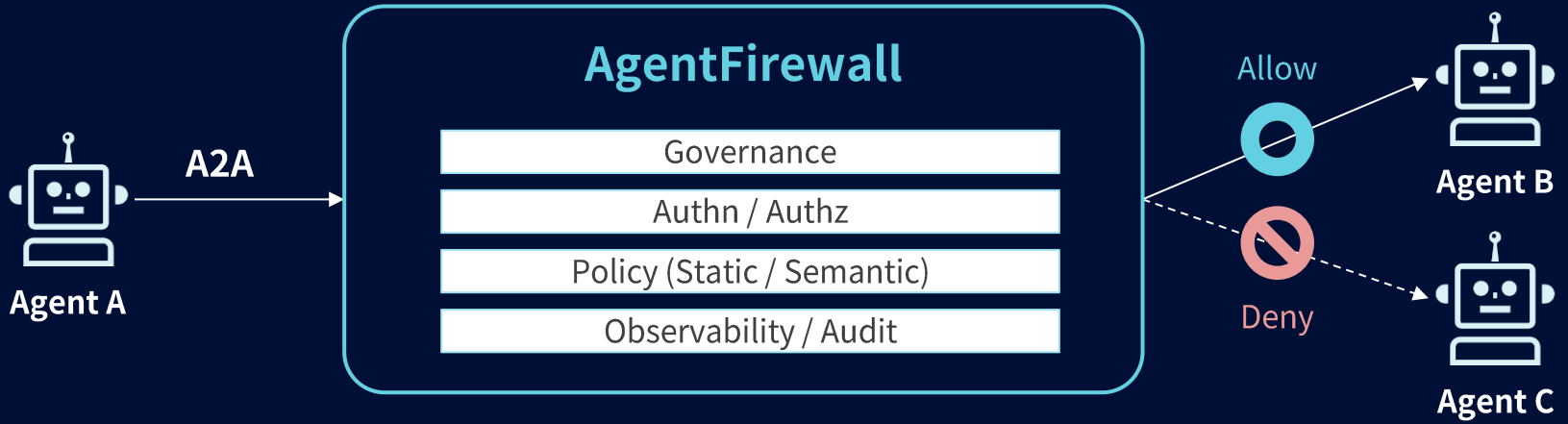
AI Agents Are Starting to Talk to Each Other



New design question:

**Who can talk to whom,
and under what authority?**

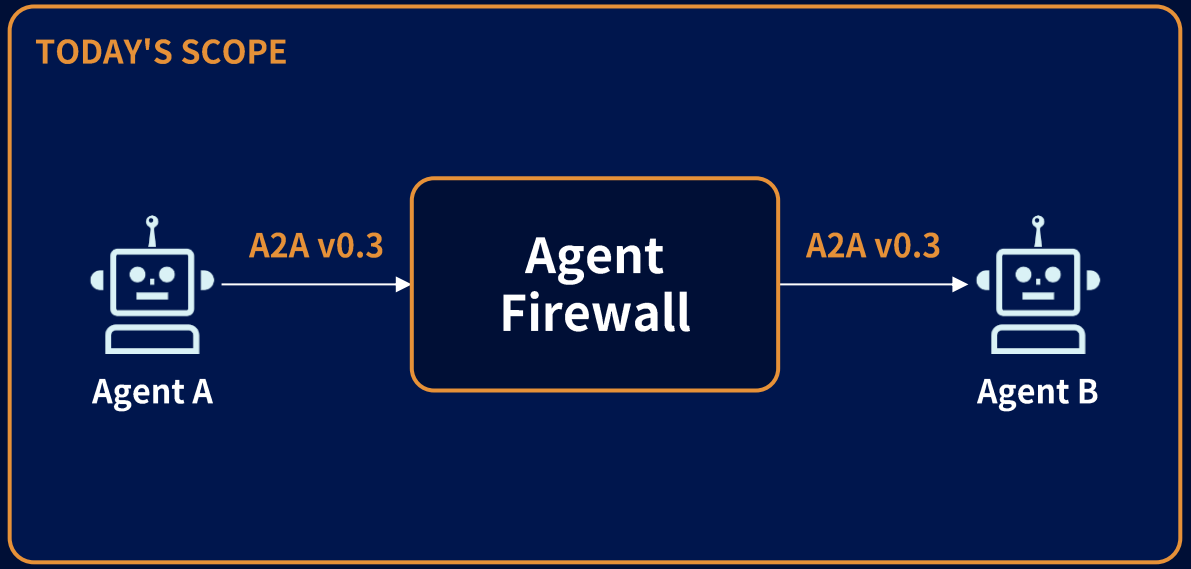
Our Starting Point: Agent Firewall



Control agent-to-agent connectivity, not just content.

https://www.softbank.jp/corp/news/press/sbkk/2025/20250724_01/

What We Actually Tested

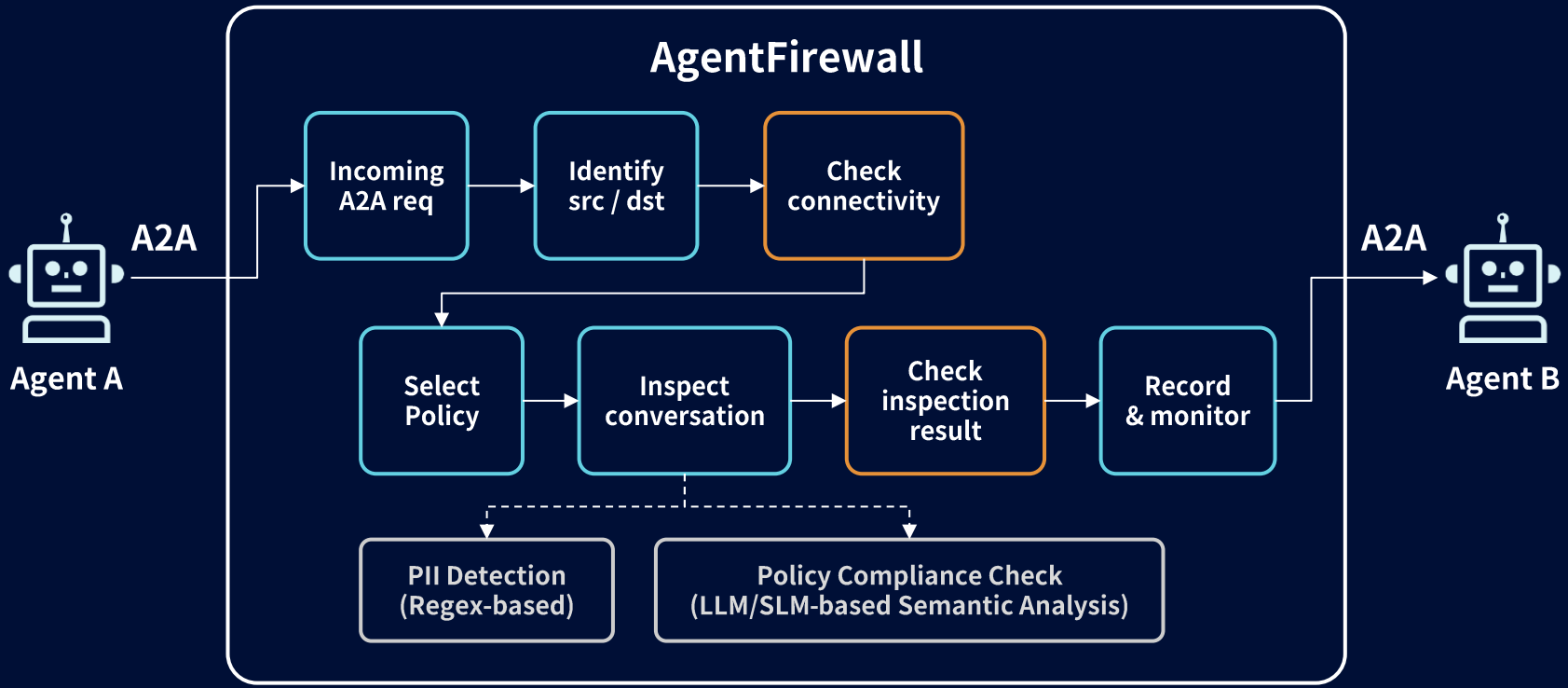


Not covered in this PoC

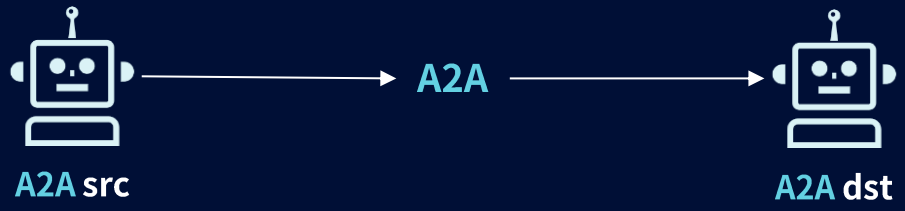
- MCP / Tool execution
- Agent → Data access
- Full AgentSecOps PF

PoC was conducted with A2A Protocol v0.3.
Architecture lessons are separated from version-specific behavior.

Our Control Model



Lesson #1 — What Should We Govern as an “Agent”?



Is an A2A endpoint the right unit of governance?

- Does speaking A2A make something an agent?
- What about agents that never speak A2A?
- What characteristic should define an agent?

Before defining the trust boundary,
we thought that define what counts as an agent.

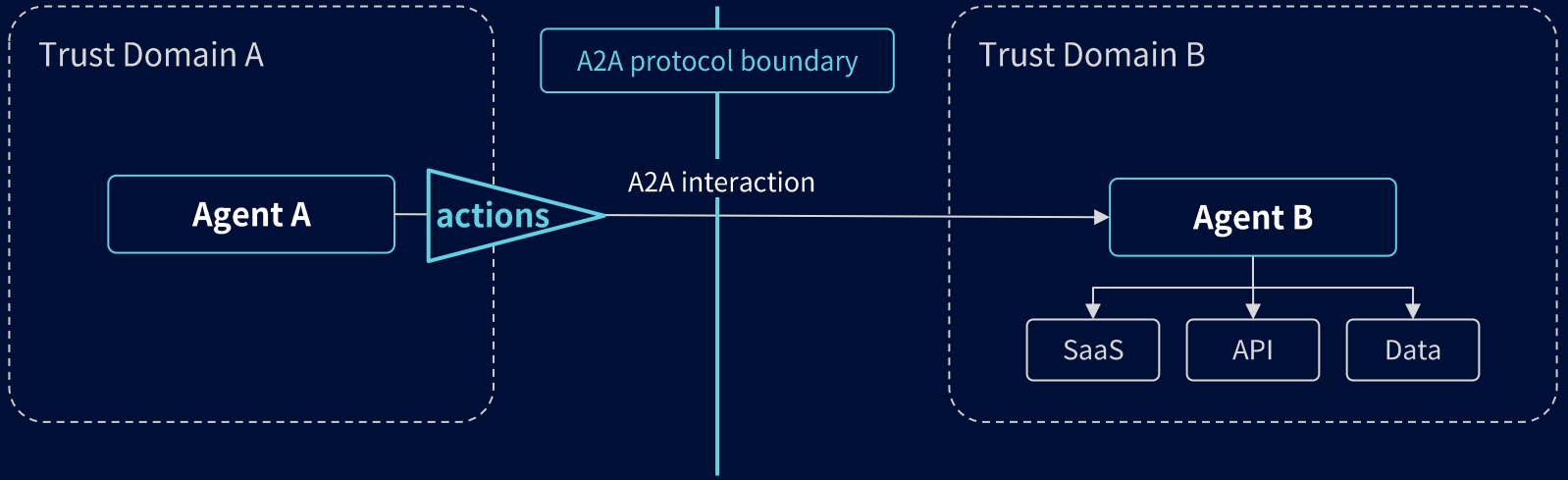
Our Definition: Decision + Action

A2A is one interaction pattern, not the definition of an agent.

A2A Agent | SaaS Agent | Local Agent

Trust Boundaries Extend Beyond A2A Boundaries

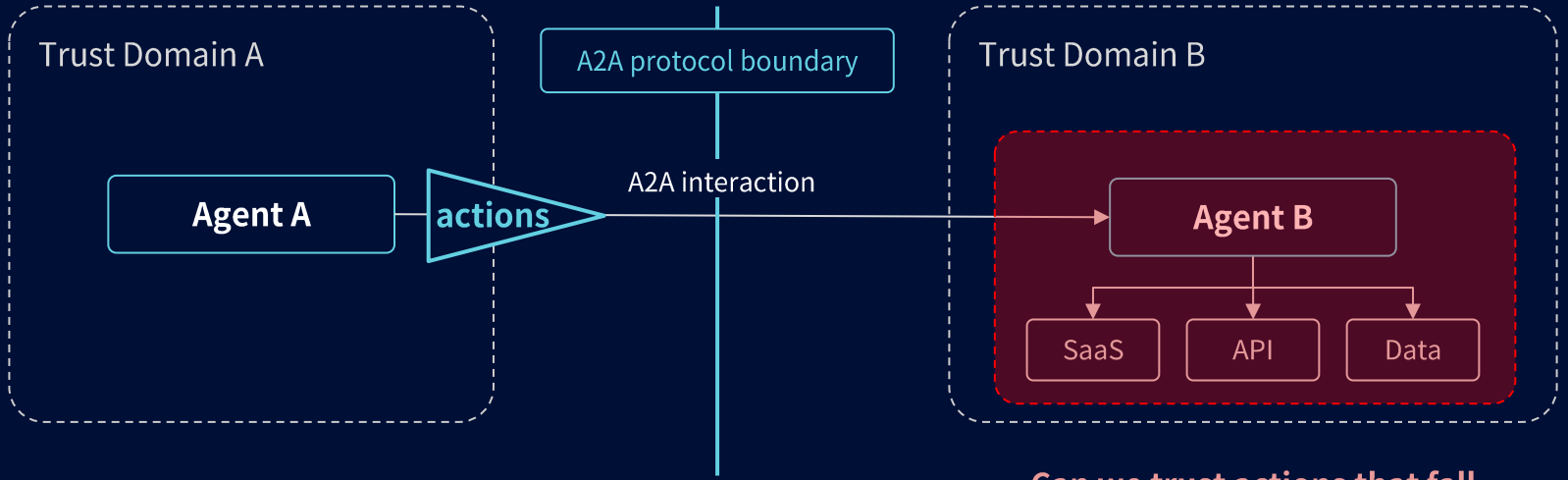
A2A tells us how agents communicate. It does not tell us where trust should end.



Protocol boundary
defined by the communication mechanism

Trust Boundaries Extend Beyond A2A Boundaries

A2A tells us how agents communicate. It does not tell us where trust should end.



Protocol boundary
defined by the communication mechanism

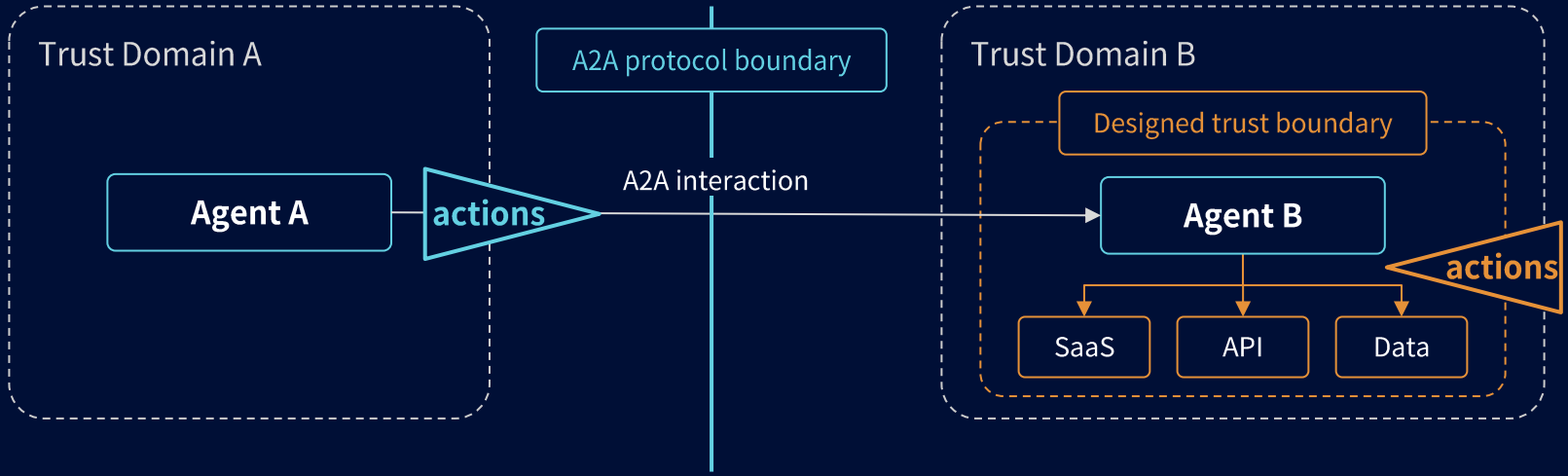
Can we trust actions that fall outside the A2A scope ?

▼

Of course NOT.

Trust Boundaries Extend Beyond A2A Boundaries

A2A tells us how agents communicate. It does not tell us where trust should end.

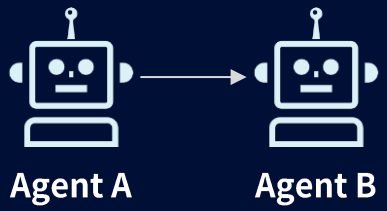


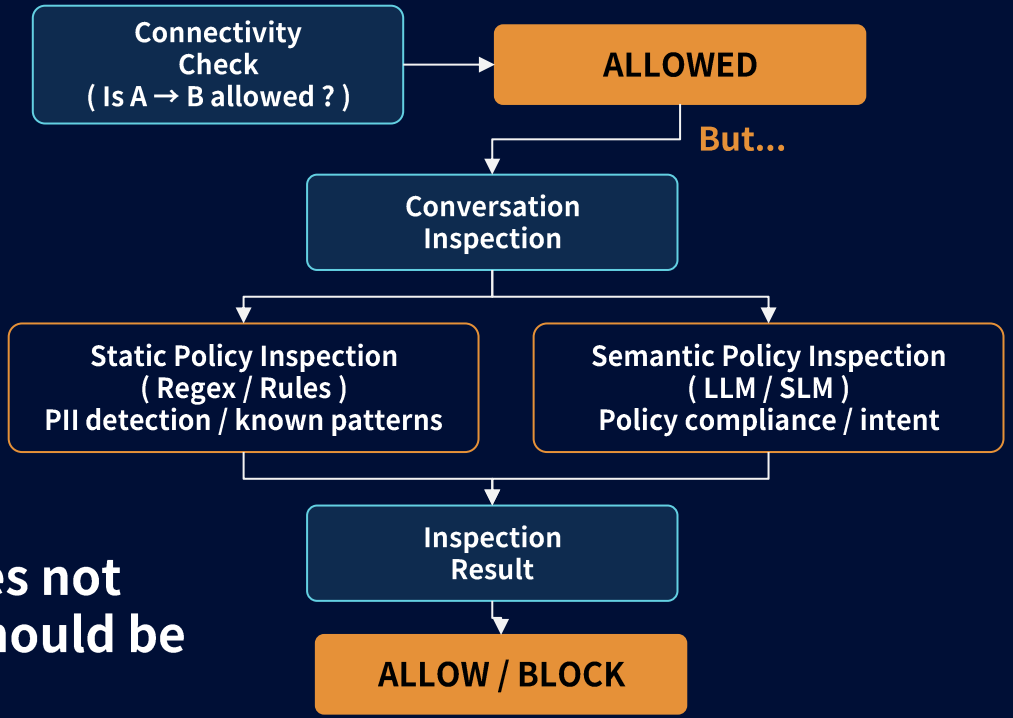
Protocol boundary
defined by the communication mechanism

Trust boundary
designed around decision, authority, and action

Lesson #2 — Policy Enforcement Goes Beyond Connectivity

Connectivity decides whether agents can talk. Inspection decides whether this conversation should proceed.





A permitted connection does not mean every conversation should be permitted.

The Need for Semantic Inspection

Lesson #2

Andrew Ng

> “LLM-as-a-Judge evals, in which you prompt an LLM to efficiently come up with ways to evaluate more open-ended outputs.” [1]

Harrison Chase (LangChain Co-founder and CEO)

> “LLM applications often produce outputs in natural language, which are difficult to judge using hard-coded rules.” [2]

> “You can't monitor agents like traditional software. Inputs are infinite, behavior is non-deterministic, and quality lives in the conversations themselves.” [3]

As AI agents become more autonomous, semantic inspection powered by models becomes essential.

[1] https://www.linkedin.com/posts/andrewyng_new-short-course-evaluating-ai-agents-evals-activity-7298024219433451521-PXow (2024)

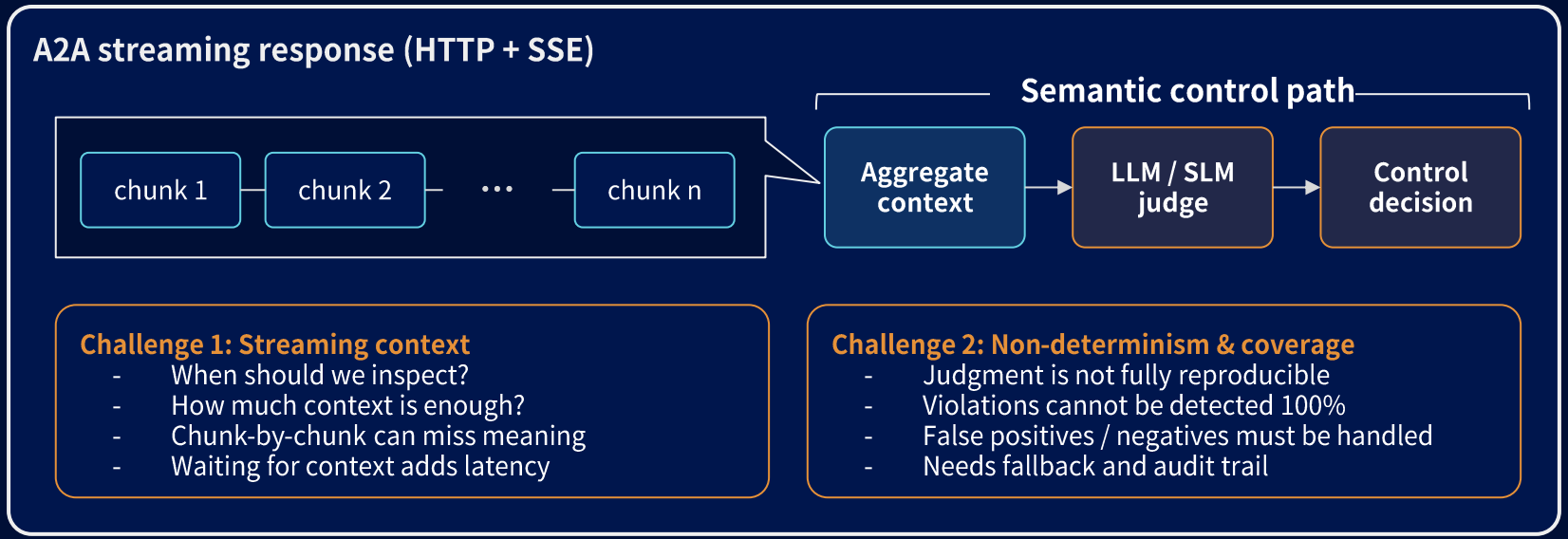
[2] https://www.langchain.com/blog/aligning-llm-as-a-judge-with-human-preferences (2024)

[3] https://www.langchain.com/blog/production-monitoring (2026)

Semantic Inspection Introduces New Challenges

Lesson #2

Zoom-in on semantic inspection: the hard part is not only judging meaning, but judging it reliably on streaming A2A traffic.

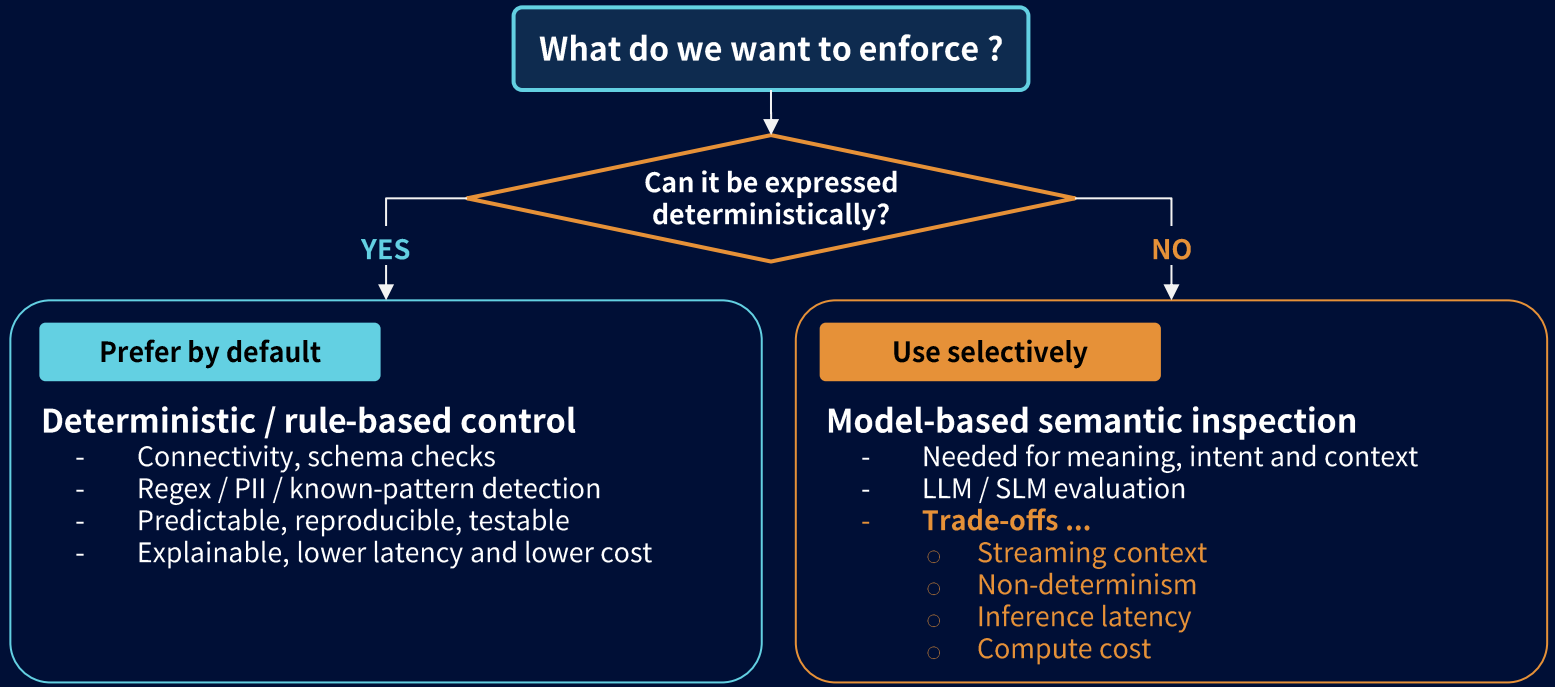


Semantic inspection is not just a model call; it is a runtime control design problem.

Lesson #3 — Stronger Inspection Comes at a Cost

Rule-based by default; model-based semantic inspection only where necessary.

Policy design principle: start from the control requirement - not from the model.



Lesson #3 — Stronger Inspection Comes at a Cost

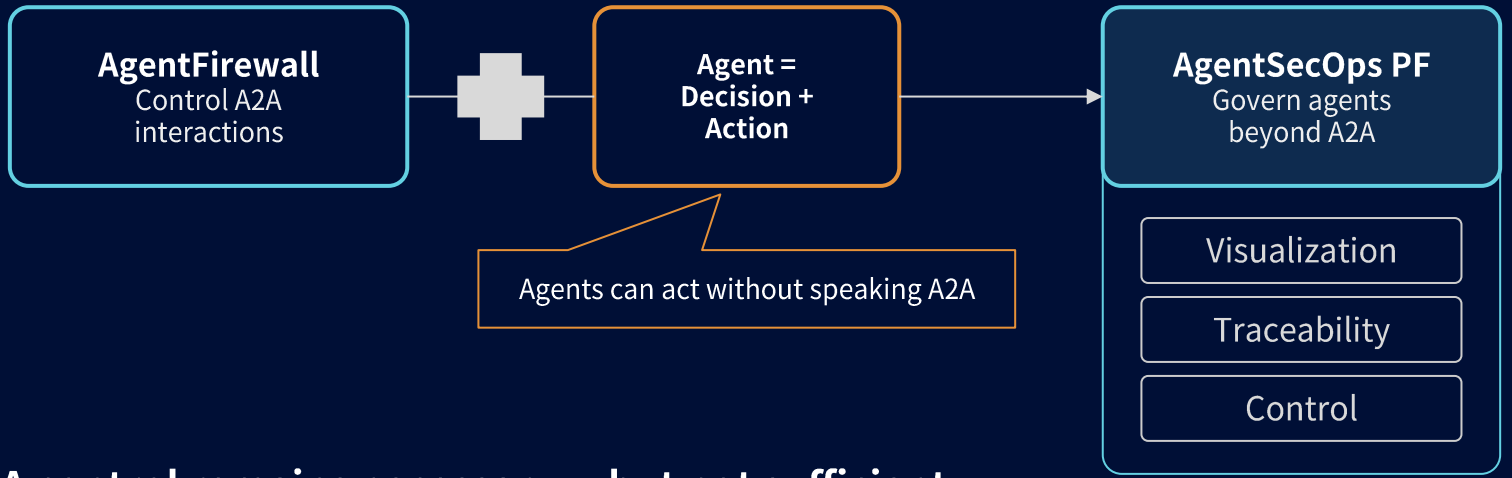
Rule-based by default; model-based semantic inspection only where necessary.

Policy design principle: start from the control requirement - not from the model.

Do NOT turn every policy intent into a prompt.
Turn it into a rule first,
unless meaning requires a model.

From AgentFirewall to AgentSecOps

The protocol stopped being the boundary of what we needed to govern.



A2A control remains necessary — but not sufficient.

The control scope expanded from communication paths to decision-making actors.

Takeaways — Three Things We Learned

01 **Define the agent before defining the trust boundary.**

02 **Trust boundaries extend beyond protocol boundaries.**

03 **Deterministic by default. Semantic where meaning matters.**

Design trust boundaries around what agents can decide and do.

What’s Next — Engaging with AAIF

SoftBank Corp. joined the Agentic AI Foundation (AAIF) this fiscal year.

We plan to bring the lessons and open questions from this work into the AAIF community, and contribute to broader discussions around trustworthy agent systems.

Thank you


SoftBank