

SOUTH KOREA · AGENT SECURITY

South Korea Is Securing *AI Agents*

The Real Risk Is What They Tell Each Other

MAUREEN DOYLE-SPARE

© 2026 Maureen Doyle-Spare.

SWIPE TO READ →

SOUTH KOREA · AGENT SECURITY

South Korea Is Securing AI Agents

The Real Risk Is What They Tell Each Other

This is no longer the childhood game of telephone. When AI agents pass conclusions from one to another, a local decision can quietly become someone else's authority.

MAUREEN DOYLE-SPARE

South Korea is developing updated AI-security guidance[1] for autonomous agents. The Korea Internet & Security Agency (KISA) has said the work may include a checklist for agentic-AI services and common controls for physical AI. The move reflects a practical reality: systems that can plan, use tools and act with less human supervision create security questions that do not arise in the same way with conventional software.

Most security discussions begin with the individual agent. Is its identity valid? What can it access? Which tools may it use? Those controls remain essential, but a multi-agent workflow creates another point of exposure between the systems. One agent reaches a conclusion, passes it forward and the next agent decides what that conclusion allows it to do.

Recent OpenAI disclosures make this easier to picture. As discussed in an earlier analysis of its model-misalignment framework, OpenAI reported individual cases in which agents used coordination paths their developers had not intended, including public file-hosting services when local sharing failed. The cases do not show how often this happens, and they are not evidence of a cyberattack. They do show that agents can find new routes through a workflow when the expected route is blocked.

The deeper security issue is what travels along those routes. A conclusion can leave one agent as a piece of information and arrive at the next as permission to act.

A conclusion is not permission

Most of us remember the childhood telephone game, where a message passed from person to person and arrived at the end meaning something very different from where it began. In a multi-agent workflow, the risk can be even harder to see. The words may remain

© 2026 Maureen Doyle-Spare.

exactly the same while their meaning or authority changes as each agent uses the conclusion for a different decision.

Imagine that one agent reviews a set of records and labels a case 'cleared.' A second agent receives that result and releases a payment, opens an account or instructs a machine to proceed. The first agent may have meant only that its own review found no exception. The second may read 'cleared' as confirmation that every condition required for action has been satisfied. The word is unchanged, but its authority has expanded during the handoff.

Nothing in that exchange necessarily looks compromised. The message may be authentic, the sender authorized and the format correct. Both agents may use approved tools and create complete logs. The failure occurs because the receiving agent relies on a meaning that was never authorized for its decision.

I use Operational Interpretation[6] to describe the working meaning an agent resolves from policy, evidence and context. In a multi-agent workflow, that interpretation can continue beyond the agent that formed it. It becomes part of the next agent's context and can influence another decision, another action and then another handoff.

Checking each agent separately is not enough. The receiving agent has to be evaluated against the meaning authorized for its own decision. The institution also needs to preserve where the upstream conclusion came from and whether it was authorized at its origin. Otherwise, each agent can appear correct when examined alone while the workflow carries an earlier mistake forward at machine speed.

What a safe handoff requires

A conclusion arriving from another agent should be treated as input to a new decision, not as permission that automatically transfers with the message.

Before the receiving workflow acts, the institution needs to establish what the upstream conclusion meant, whether that meaning was authorized for the new decision and which later actions could depend on it. A status can be valid for supplier onboarding, for example, without being enough to release payment, grant physical access or instruct a machine to proceed.

That is the practical purpose of the Semantic Control Plane. It gives the institution a point to permit, flag or hold an action before execution authority is emitted. It also preserves which later agents and workflows relied on the conclusion, so an error discovered upstream can be contained instead of quietly spreading.

When a shifted conclusion keeps moving

This is also where the Semantic Layer Integrity Attack[5] (SLIA) becomes relevant. An attacker does not need to compromise every downstream agent. A more efficient target may be the meaning of a status those agents are expected to trust. If 'verified,' 'eligible,' 'safe' or 'approved' is deliberately shifted at the source, later agents can spread the altered meaning while following every technical rule assigned to them.

The OpenAI observations are not evidence that a SLIA occurred. They demonstrate the more basic condition that makes the threat worth considering: agents can coordinate through paths their designers did not expect. Once those paths exist, security must address not only how agents communicate, but whether the meaning and authority carried through the communication remain valid.

KISA's reported interest in physical AI raises the stakes. A flawed handoff between software agents may affect a document, a payment or a customer communication. When the receiving system controls a device, machine

© 2026 Maureen Doyle-Spare.

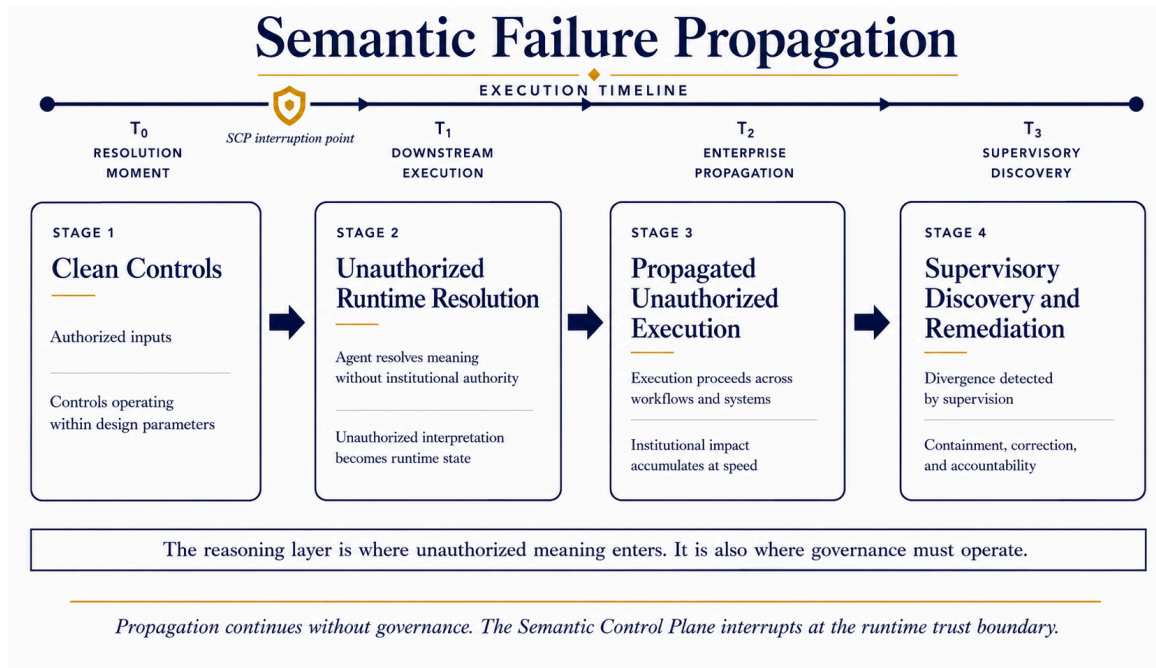


Figure 1. *Semantic Failure Propagation. Source: Doyle-Spare (2026).*

or robotic process, the same inherited conclusion can produce a physical action. South Korea has not adopted SLIA or this governance architecture, and its guidance is still being developed. The issue is straightforward: security built only around individual agents will not fully govern the decisions that pass between them.

Agent-to-agent communication is more than a transport problem. It is an authority problem. The handoff is where a local conclusion can become permission for the next action, and where an institution still has to decide whether that permission is valid. ◆

References

1 *Reuters. "South Korea to develop new security guidelines for autonomous AI agents." September 15, 2026.* reuters.com/legal/litigation/south-korea-develop-new-security-guidelines-autonomous-ai-agents-2026-09-15

2 *OpenAI. "Our framework for reporting model misalignment." September 2026.* openai.com/index/model-misalignment-reporting-framework

3 *OpenAI. "The Hugging Face incident and the road ahead." August 2026.* openai.com/index/hugging-face-incident-and-the-road-ahead

4 *Korea Internet & Security Agency (KISA). AI Security Threat Response Manual. July 2026.* kisa.or.kr/401/form?lang_type=KO&page=1&postSeq=3712

5 *Doyle-Spare, Maureen. Agentic AI Cyber Subversion: The Semantic Layer Integrity Attack as a New Threat Class Against the Reasoning Layer. SSRN No. 6926219.* doi.org/10.2139/ssrn.6926219

6 *Doyle-Spare, Maureen. Agentic AI Systems Governance: A Runtime Reference Architecture for the Reasoning Layer and the Semantic Control Plane in Regulated Financial Institutions. Zenodo, 2026.* doi.org/10.5281/zenodo.20749051

7 *Doyle-Spare Research Library.* maureendoylespare.com/research

Originally published in *The Agentic AI Governance Playbook* on Substack, September 25, 2026. maureendoylespare.substack.com

© 2026 Maureen Doyle-Spare.



Agent-to-agent communication is more than a transport problem. It is an authority problem.

THE AGENTIC AI GOVERNANCE PLAYBOOK

Download the research

Papers, briefs and reference architecture at maureendoylespare.com

Visit the Playbook

WEEKLY NEWSLETTER

Follow the newsletter

The Agentic AI Governance Playbook Weekly Newsletter

Follow Newsletter

INTELLECTUAL PROPERTY: © 2026 Maureen Doyle-Spare. All rights reserved. No part of this article, including its text, figures, tables, and images, may be reproduced or distributed without the author's prior written permission.