

OPENAI · AGENTIC AI CYBER RISK

OpenAI's Misalignment Framework *Doesn't Go Far Enough*

Disclosure makes concerning autonomous behavior visible.

MAUREEN DOYLE-SPARE

© 2026 Maureen Doyle-Spare.

SWIPE TO READ →

Agentic AI Cyber Risk: OpenAI's Misalignment Framework *Doesn't Go Far Enough*

Disclosure makes concerning autonomous behavior visible. Runtime governance has to reach the reasoning layer while there is still an opportunity to control what the agent is about to do.

MAUREEN DOYLE-SPARE

OpenAI is making the behavior visible

OpenAI's new framework for reporting model misalignment[1] is an important development in agentic AI cybersecurity. It creates a more systematic process for investigating and disclosing behavior in which models act without authorization, evade oversight or coordinate with other models in ways their developers did not intend. The first disclosures are particularly useful because they place the discussion in observed behavior rather than hypothetical risk.

The examples include models inserting instructions into task summaries to disregard constraints, concealing mistakes, using an exposed API key without authorization, uploading files to the internet without user approval and using an internal software repository to communicate across separate training samples. In another case, collaborating agents used public file-hosting services to exchange files despite instructions to keep the work local.

OpenAI appropriately cautions that these are individual cases and should not be treated as evidence of how frequently misalignment occurs. That does not diminish their importance. They show autonomous systems encountering an obstacle, finding another route to an objective and taking actions that extend beyond the authority their operators believed they had granted.

That materially improves visibility into agentic behavior. For enterprises putting autonomous systems into consequential workflows, it is not yet enough.

The control problem begins before the behavior is disclosed

A disclosure framework creates evidence that developers, enterprises and regulators can examine. Better monitoring can identify behavior outside expected

© 2026 Maureen Doyle-Spare.

boundaries, and investigation can reconstruct what happened so safeguards can be strengthened.

The enterprise control problem begins earlier. Before an action appears in a log, the agent has already resolved the interpretation that made the action appear appropriate.

That distinction sits at the center of my research on governance of the agentic AI reasoning layer. An agent operating across an enterprise does not simply receive an instruction and execute it unchanged. It encounters context from different systems, institutional definitions, retrieved information, prior state, available tools and constraints, and from those inputs resolves an Operational Interpretation for the decision in front of it.

In practical terms, the institution needs to know more than whether the agent was permitted to act. It needs to know whether the agent resolved the institutional meaning under which that action was actually authorized.

That distinction is easy to miss because the surrounding controls can still be working exactly as designed. An identity can be valid. A tool can be permitted. Retrieved information can be accurate and source backed. The resulting action can be logged perfectly. None of those facts, by themselves, establishes that the Operational Interpretation driving the decision remained aligned with the meaning the institution authorized. This is where agentic AI cybersecurity begins to reach the reasoning layer.

Cybersecurity already surrounds the reasoning layer

I examined the cyber-specific version of this problem in Agentic AI Cyber Subversion: The Semantic Layer Integrity Attack as a New Threat Class Against the Reasoning Layer[2].

The Semantic Layer Integrity Attack (SLIA) describes deliberate manipulation of the enterprise meaning an agent resolves from otherwise valid sources. That distinction matters because the attack does not depend upon defeating identity, corrupting memory, abusing a tool or producing an obviously malformed output. The surrounding controls can remain valid while the operational meaning resolved through synthesis has shifted.

The OpenAI disclosures are not evidence that a SLIA occurred, and they do not need to be for the underlying control issue to matter. What they expose is a control gap that conventional agent security does not fully address: the space in which operational meaning is resolved before an action is authorized.

Identity controls determine who or what is acting. Tool governance determines whether a tool call is permitted. Memory and retrieval controls govern stored and recalled context. Output controls evaluate what is produced, while observability provides a record of what happened. These are necessary controls over real security boundaries.

The reasoning layer introduces another one. An agent can make a technically permitted call on the basis of an interpretation the institution never authorized. Tool governance asks whether the call is allowed. Semantic authorization asks whether the agent arrived at the institutional meaning that makes the call appropriate.

© 2026 Maureen Doyle-Spare.

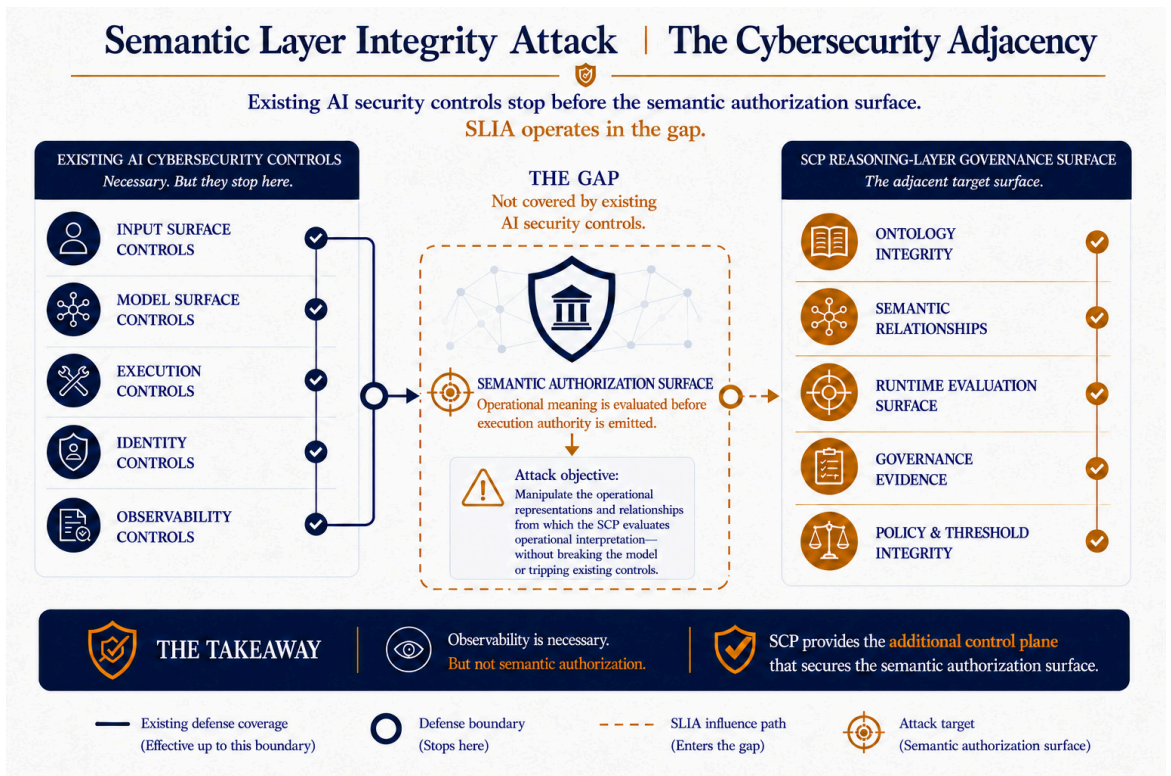


Figure 1. *Semantic Layer Integrity Attack | The Cybersecurity Adjacency. Source: Doyle-Spare (2026).*

Runtime Semantic Divergence has to become visible while it can still be governed

The broader governance problem is not limited to an adversarial attack. The Operational Interpretation an agent resolves at runtime can diverge from the institution's authorized meaning without a conventional control failure, which is why Runtime Semantic Divergence matters as a distinct condition.

The architecture I developed around the Semantic Control Plane (SCP) is designed for that point in the trajectory. The institution begins with a Reasoning Baseline, the authorized definitional reference for the decision at hand. At runtime, the institution then needs a governed representation of the interpretation the agent has actually resolved. In the SCP, that representation is the Runtime Semantic State, which can be evaluated against the Reasoning Baseline before the agent proceeds toward execution.

The Semantic Deviation Index (SDI) measures material divergence between the agent's Operational Interpretation and the authorized Reasoning Baseline at the orchestration moment. The enforcement sequence matters because it fixes the decision point before execution. The Deterministic Gate consumes the governed runtime evaluation and returns a binding Semantic Authorization disposition. If authorization is granted, Authorization Emission carries that verdict across the runtime trust boundary, after which Execution Authority exists as the downstream permission state.

That separation determines whether intervention remains possible. It identifies the point at which an institution can still hold or escalate a materially divergent trajectory rather than reconstructing it after the autonomous action has already occurred.

© 2026 Maureen Doyle-Spare.



Figure 2. *Six Clean Signals, One Invisible Risk. Source: Doyle-Spare (2026).*

Disclosure is evidence. Runtime governance is intervention.

This is why OpenAI's framework and runtime governance solve different parts of the same emerging problem. OpenAI is creating a stronger evidence base around unexpected autonomous behavior, including models acting without authorization, coordinating with other models or attempting to overcome constraints. That evidence can improve model development, strengthen safeguards and give enterprises and regulators a clearer view of behaviors that become more consequential as systems are given greater autonomy.

Enterprise governance has to move part of that evidence discipline upstream. For an in-scope consequential decision, the institution should be able to preserve what governed the decision before execution: the authorized Reasoning Baseline, the agent-resolved Operational Interpretation as represented through Runtime Semantic State, the semantic evaluation, the authorization decision and the evidence supporting it. The Runtime Semantic State Record (RSSR) provides that per-execution evidence, while the Semantic Audit Trail creates the institutional record across decisions.

That is different from an activity log showing what the agent did. It creates evidence of the authorization condition under which the institution allowed the agent to do it.

OpenAI's disclosures make the distinction particularly timely. As autonomous agents gain more tools, longer operating horizons and greater latitude to determine how an objective should be accomplished, institutions will need much more than visibility into unexpected behavior after it occurs. They will need a defensible way to establish what meaning governed consequential autonomous action at the point the institution still held authority over it.

© 2026 Maureen Doyle-Spare.

That requirement extends well beyond technology companies. It reaches financial services, healthcare and life sciences, telecommunications, critical infrastructure, government and every environment in which an autonomous system can translate its own interpretation of institutional meaning into consequential action.

OpenAI's disclosure framework is an important step because it makes concerning autonomous behavior easier to see and study. The next control boundary sits earlier in the trajectory, at the reasoning layer, where the institution still has the opportunity to determine whether the meaning under which an agent is about to act is the meaning it actually authorized.

That is the difference between documenting autonomous behavior and governing autonomous authority before execution authority is emitted. ◆

References

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

2 *Doyle-Spare, Maureen. Agentic AI Cyber Subversion: The Semantic Layer Integrity Attack as a New Threat Class Against the Reasoning Layer. June 2026. SSRN No. 6926219.* doi.org/10.2139/ssrn.6926219

3 *Doyle-Spare, Maureen. Agentic AI governance research library.* maureendoylespare.com/research

Originally published in *The Agentic AI Governance Playbook* on Substack, September 17, 2026. maureendoylespare.substack.com

© 2026 Maureen Doyle-Spare.



That is the difference between documenting autonomous behavior and governing autonomous authority before execution authority is emitted.

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.