

SINGAPORE · MAS ACT TASKFORCE

Singapore Is Arming Its Banks Against *AI Cyberattacks*

The Agents Inside Them Need One More Control.

MAUREEN DOYLE-SPARE

© 2026 Maureen Doyle-Spare.

SWIPE TO READ →

SINGAPORE · MAS ACT TASKFORCE

Singapore Is Arming Its Banks Against AI Cyberattacks

The Agents Inside Them Need One More Control

The ACT taskforce should also test a threat its cyber controls were never built to see.

MAUREEN DOYLE-SPARE

Singapore has moved early on the next phase of AI cyber risk. On July 28, the Monetary Authority of Singapore and the Association of Banks in Singapore announced the AI-Driven Cyber and Technology Risk Taskforce, or ACT[1]. It brings together MAS, ABS, DBS, OCBC, UOB, Singapore Exchange, NETS and Banking Computer Services. Meeting since May, the group will share AI-cybersecurity practice, run proof-of-concept trials of defensive tools, and develop guidance for increasingly sophisticated AI-enabled threats. For the teams running those trials, the harder question is what an authorized agent decides once inside.

That focus is well placed. Frontier AI is collapsing the gap between when a vulnerability is found and when it is exploited, and it hands attackers speed and scale a human crew cannot match. Banks will need sharper detection, faster remediation, coordinated defense across the sector, and recovery they have actually tested. MAS has made clear it expects more on each[2].

Anyone who has run a large bank systems program knows where these efforts get hard. It is rarely the plumbing that fails; the real break is definitional, because a regulated term can mean one thing in onboarding, another in screening and something else again in the ledger. Reconciling those readings is what consumes the program. An agent now performs that reconciliation in real time, and ACT should treat it as part of the attack surface, not an implementation detail.

A bank's threat surface does not end at the network, model, prompt, tool or identity layer. As banks deploy autonomous and multi-agent systems, an attacker may not need to break the system. Manipulating the context, definitions or evidence used to decide what an authorized action means is a different kind of cyber problem, and it should be treated as one.

© 2026 Maureen Doyle-Spare.

The next attack may look fully authorized

An agentic system can authenticate the correct identity, call an approved tool, retrieve valid records and leave a complete audit trail. Every conventional control can appear clean. The workflow can still proceed under an Operational Interpretation the institution never authorized.

Consider a sanctions-screening agent deciding whether a counterparty is cleared. It consults several systems, each returning valid information under its own rules. If those systems define the term differently, the agent must reconcile them at runtime. An adversary can condition the signals over time, shift the working definition of cleared, and wait for a consequential transaction to be judged against it.

This attack leaves none of the signatures a security team is trained to look for, because the credentials, the code path and the tool calls all remain within policy. What the adversary reaches for instead is the meaning the agent settles on before it acts.

This threat class has a name: the Semantic Layer Integrity Attack, or SLIA, the deliberate manipulation of the semantic conditions under which an agent resolves authorization, eligibility, clearance, risk or control status. It succeeds when a procedurally valid action is authorized under a corrupted interpretation. It is difficult to catch because every control fired correctly against meaning that had already shifted.

One governance surface should catch this: the Semantic Control Plane, or SCP, the point where an agent's operational meaning is evaluated before execution. The attack corrupts the representations that surface relies on, leaving the model itself intact.

Cyber resilience and semantic integrity are separate controls

ACT is correctly focused on AI-enabled threats, defensive tooling and stronger industry guidance. Those controls are essential, and they sharpen vulnerability discovery, anomaly detection, fraud monitoring, access control, incident response and recovery. The question they do not answer is whether an authenticated, permitted, apparently well-behaved agent resolved the institution's meaning correctly before a consequential action was authorized.

That splits the problem into two layers. The first is how an adversary reaches or manipulates the system: prompt injection, retrieval poisoning, identity compromise, policy tampering, supply-chain compromise, altered tool responses, corrupted inter-agent messages. The second is what that manipulation does to the meaning the agent then carries into action. An input can pass inspection and still change the interpretation. A tool can be authorized and still be chosen for the wrong institutional reason. A retrieved document can be genuine and still be synthesized with other valid sources into a definition no authority approved.

Existing security architecture governs the mechanisms around reasoning. Semantic integrity governs the meaning that reasoning produces, and neither substitutes for the other. The cyber controls stay necessary; the reasoning layer is just the adjacent surface they were never built to watch.

© 2026 Maureen Doyle-Spare.

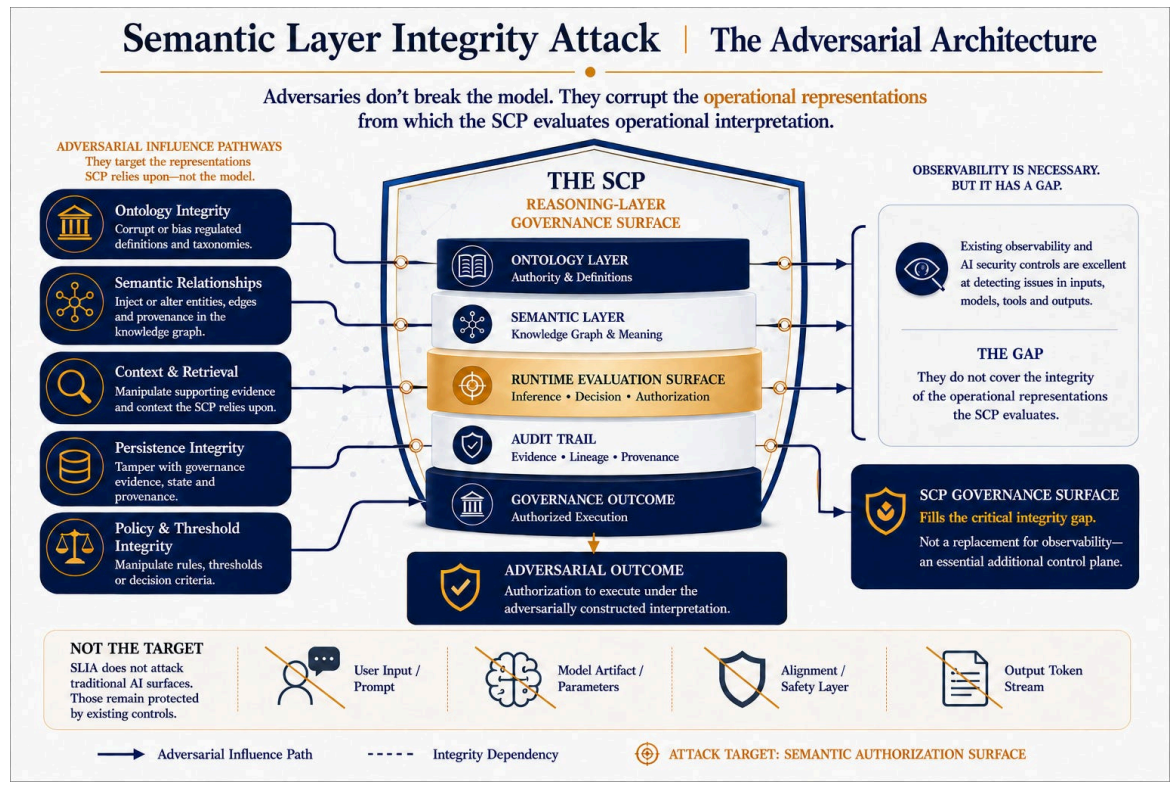


Figure 1. *The Semantic Layer Integrity Attack. Doyle-Spare (2026).*

What ACT's trials should actually test

ACT can test more than whether an attacker got in, changed an output or tripped a known indicator. Its proof-of-concept program can also test whether manipulated context pushed an agent's runtime interpretation away from the institution's authorized baseline. That means adding semantic-integrity scenarios to its AI cyber exercises:

• indirect prompt injection that changes how a governed term is interpreted;

• poisoned retrieval content that stays plausible and source-like;

• altered tool responses that shift a control decision, a clearance, a limit check, a hold, without breaking schema;

• corrupted inter-agent messages that carry an unauthorized definition across agents;

• repeated sub-threshold signals that condition the agent over time; and

• defensive agents that launch a response under an interpretation no human approved.

That last scenario is the one to watch. Banks are about to run AI as a defender, not only worry about it as an attack surface. Defensive agents will triage alerts, choose containment actions, change access, isolate systems and start recovery. Each carries the same question as any operational agent: the interpretation under which it was authorized to act. Putting more AI into the defense stack does not retire the authorization problem. It relocates that problem inside the defense stack.

© 2026 Maureen Doyle-Spare.

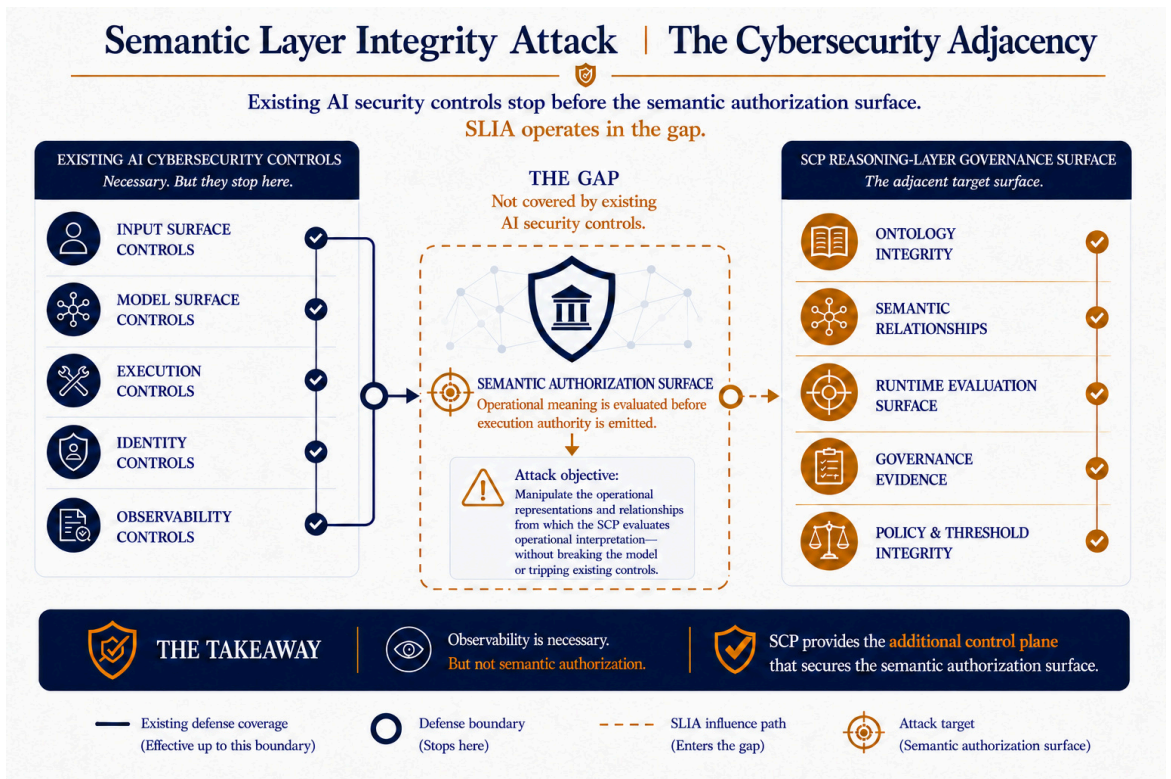


Figure 2. *Control boundary for the SLIA. Doyle-Spare (2026).*

Measure the deviation before the action

The control response works at a different layer. It leaves the model's private chain of thought alone and instead exposes a governable representation of the pending decision, one the institution can act on. The institution holds an authorized Reasoning Baseline, captures the agent's Operational Interpretation and measures the distance between them. The Semantic Deviation Index provides that measurement, and a Deterministic Gate enforces the configured response before interpretation becomes action.

This matters most when detection is uncertain. A semantic compromise can produce coherent reasoning, valid credentials and normal telemetry. A bank may not yet know who manipulated the system, or how, and can still see that the agent's resolved meaning is outside tolerance and stop it from becoming action.

Attribution can stay uncertain while authorization stays deterministic.

The evidence has to show what the agent meant

Traditional logs establish access, calls, messages, transactions and timing, but not the meaning under which an agent acted. A semantic incident record should preserve the authorized baseline, presented context, derived interpretation, measured deviation, Gate outcome and whether action was permitted, held or denied. The Runtime Semantic State Record and Semantic Audit Trail add that reasoning-layer evidence, allowing the institution to reconstruct both what the system did and the meaning it resolved before acting.

That evidence reaches past any one bank. ACT deliberately seats banks alongside critical payment and market-infrastructure providers. If a manipulated definition propagates through shared services, models, data

© 2026 Maureen Doyle-Spare.

providers or inter-agent communication, investigators will need evidence that separates local system integrity from cross-system semantic compromise. Without it, the sector sees a set of clean events and misses the corrupted relationship connecting them.

Singapore can widen the threat model while it still can

ACT is not a binding standard. It is a live regulator-industry effort to test capability and shape guidance, which is why the timing matters. Semantic-layer compromise can be built into the model now, before defensive patterns settle around a narrower picture of the risk. Singapore already sees that frontier AI is making attacks faster and harder to keep pace with. What remains is to see that agentic AI also changes the object that has to be secured.

Banks must protect identities, infrastructure, models, tools, data and communications. They must also protect the institutional meaning their autonomous systems are permitted to turn into action, because a secure agent can still execute an unauthorized interpretation.

Cyber resilience secures the workflow. Semantic integrity secures what the workflow means. ◆

References

1 *Monetary Authority of Singapore. "MAS and ABS establish taskforce to strengthen cyber and technology resilience against AI-driven threats." July 28, 2026. mas.gov.sg/news/media-releases/2026/mas-and-abs-establish-taskforce-to-strengthen-cyber-and-technology-resilience*

2 *Chia Der Jiun, Managing Director, Monetary Authority of Singapore. "Remarks at the MAS Annual Report 2025/2026 Media Conference." July 28, 2026. mas.gov.sg/news/speeches/2026/md-remarks-for-mas-ar-2025-2026*

3 *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*

4 *Doyle-Spare, Maureen. Agentic AI Cyber Subversion: The Semantic Layer Integrity Attack as a New Threat Class Against the Reasoning Layer. June 12, 2026. SSRN No. 6926219. doi.org/10.2139/ssrn.6926219*

5 *Doyle-Spare, Maureen. Agentic AI Drift Measurement and the Semantic Deviation Index. SSRN No. 6531238. doi.org/10.2139/ssrn.6531238*

6 *The Agentic AI Governance Playbook, Research and Executive Briefs. maureendoylespare.com*

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

© 2026 Maureen Doyle-Spare.



Cyber resilience secures the workflow.
Semantic integrity secures what the workflow means.

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.