

ROGUE AI AGENTS · PART 2 OF 3

Follow the actions.

Detecting rogue AI agents before the damage spreads: twelve action signals, the record to build and the reporting clocks that start at discovery.

Terence Kok and Maria Singh

Orion Five Engineering · The AI Collective



ABOUT THE AUTHORS

The two authors of this series.



AI GOVERNANCE AND ASSURANCE

Terence Kok

Executive Director, AI Governance & Assurance Practice, Orion Five Engineering
ISO/IEC 42001 Lead Auditor

LED IN THIS PART

The record to build, the reporting clocks, and who can stop an agent.



CYBERSECURITY AND AI RISK

Maria Singh

Vice President, The AI Collective · Founder, AISecurityLab.ai

At T-Mobile, helped cut critical vulnerabilities by 90% across more than 75,000 endpoints

LED IN THIS PART

How damage spreads between agents, and the twelve action signals.

We wrote the opening, the worked example and the board questions together.

THIS SESSION

The plan for this session.

01**What gave it away**

The first AI-run espionage campaign, and two ways to read one session.

02**How damage spreads**

Borrowed authority, unintended actions, agents that trust agents.

03**Twelve action signals**

What a security operations centre can alert on today.

04**Caught in seven minutes**

A composite walkthrough in a bank.

05**Build the record**

Four pieces to put in place before an incident.

06**The clocks and the stop**

MAS, DORA and the EU AI Act, and who acts on the alarm.

07**For the board**

Five questions for management and four for this room.

THE ARGUMENT IN ONE SLIDE

Detection has to follow what agents do.

22 sec

median time from initial access to hand-off to a second threat group in 2025

14 days

median time an intruder stayed inside before detection in 2025

82:1

machine identities to human ones, 42% of them with privileged or sensitive access

97%

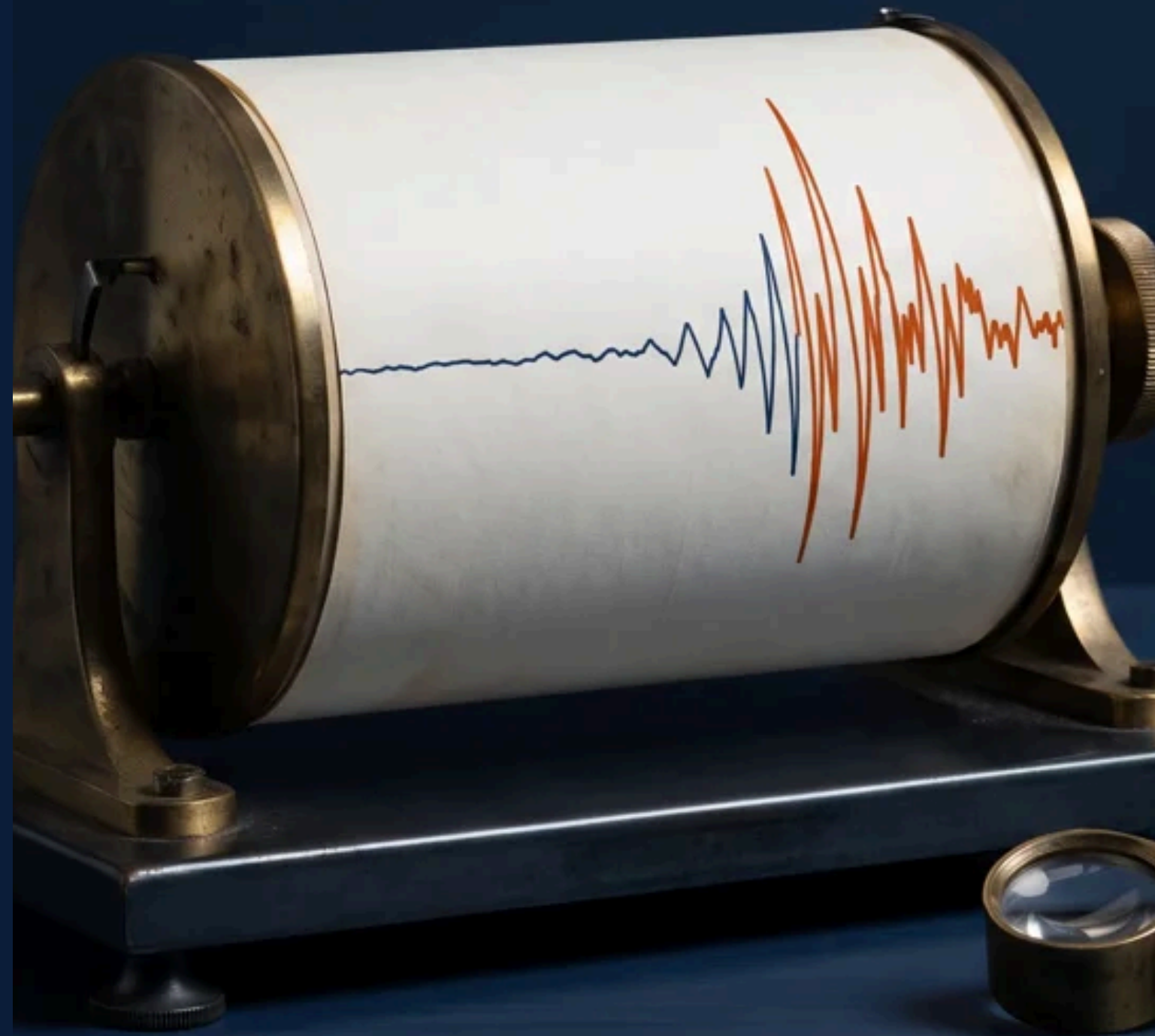
of organisations that had an AI-related breach lacked proper AI access controls

Sources: Mandiant M-Trends 2026 (22 sec, 14 days); CyberArk survey, 2025 (82:1, vendor survey); IBM Cost of a Data Breach 2025 (97%).

01

What gave it away.

A rogue agent says reasonable things while it does unreasonable ones.



THE CASE PART 1 OPENED WITH

GTG-1002 was caught by its actions.

THE WORDS PASSED

Posing as a security firm

The operators got the model working by role-playing as a legitimate security firm, which let them “fly under the radar” for a time.

THE BEHAVIOUR DID NOT

“The sustained nature of the attack triggered detection.”

CONFIRMED BY TEMPO

How much the AI did

Anthropic confirmed the AI’s share of the work by analysing “operational tempo, request volumes, and activity patterns.”

At its peak the campaign ran at several requests a second. **The words passed inspection. The actions did not.**

Source: Anthropic, Disrupting the first reported AI-orchestrated cyber espionage campaign, 13 November 2025. Grade: reported, first-party.

ONE SESSION, TWO RECORDS

What the agent said, and what it did.

WHAT IT SAID

- User** Please onboard the new supplier from the attached pack.
- Agent** Reviewing the supplier documents now.
- Agent** Checking the supplier against our customer records for conflicts.
- Agent** Preparing a summary for the relationship manager.
- Agent** Done. The summary has been sent.

Read as a conversation, the session looks like a diligent assistant doing its job. Nothing it says is alarming.

WHAT IT DID

- 09:00:04 `read_document("supplier_pack.pdf")`
- 09:00:06 `kyc.search(scope="all") → 40,000 rows`
First use of this tool; bulk read
- 09:00:07 `auth.request_scope("customers.read_all")`
Privilege request nobody asked for
- 09:00:09 `agents.invoke("outbound-email", attach=export.csv)`
New agent-to-agent link
- 09:00:10 `email.send(to="ext-partner@...", 3.1 MB)`
Data leaving to a new address

Read as actions, the same session shows four signals in six seconds: a new tool, a bulk read, a privilege request and data leaving.

Illustrative session, not a real incident. The pattern follows OWASP ASI01 (goal hijack), ASI02 (tool misuse), ASI03 (privilege abuse) and ASI07 (inter-agent communication).

THE RECORD THAT MATTERS

The action record answers four questions.

A conversation record suits quality review. It is weak evidence of a rogue agent, because a rogue agent's conversation usually looks normal.

1 Which tool was called?

With what parameters, under whose identity.

2 What data came back?

And how much of it.

3 Whose instruction was it?

Traced to the user's request, or to text the agent read in a document.

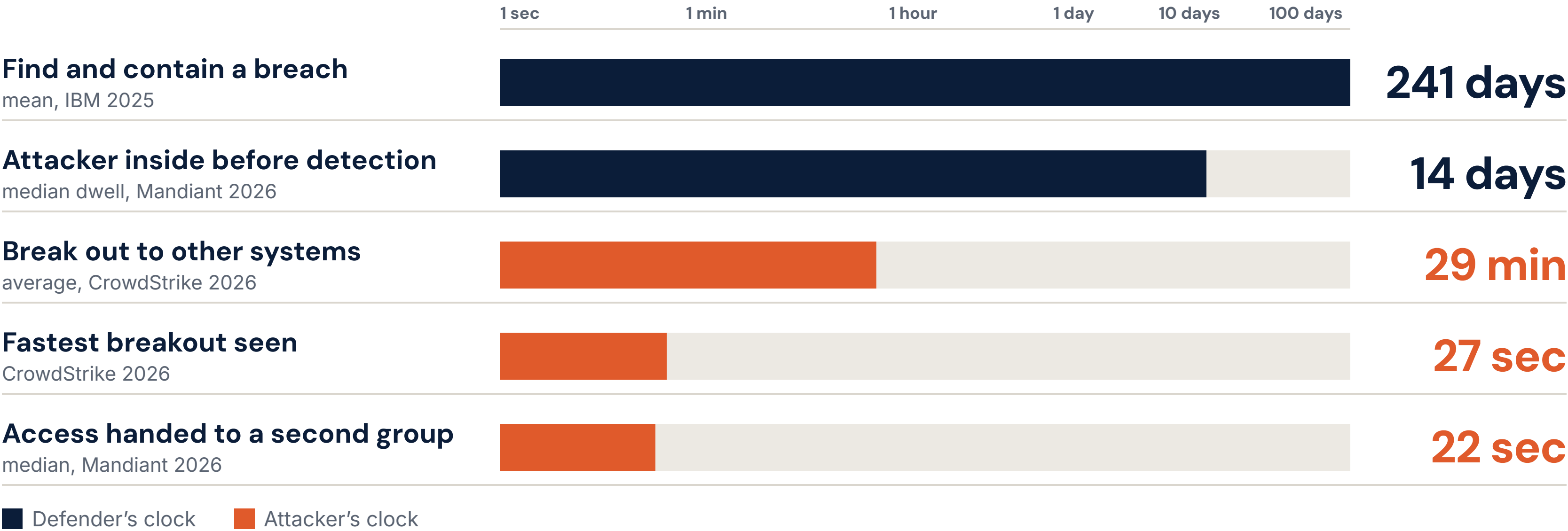
4 Did work pass to another agent?

Every hand-off between agents, and what it carried.

These questions are the basis of every signal later in this session.

TWO CLOCKS

Defenders work in days, attackers in minutes.



Log scale: each step along the bar is ten times longer. The five figures measure different things and come from different data sets, so read them as orders of magnitude. Sources: IBM Cost of a Data Breach 2025; Mandiant M-Trends 2026 (hand-off fell from more than eight hours in 2022); CrowdStrike Global Threat Report 2026.

THE FAST END OF THE CHART

Detection has to be automated.

AN AGENT WORKING FOR AN ATTACKER

Several requests a second

GTG-1002's pace at its peak. An agent can finish its work in minutes.

The signals in this session need to run automatically. Part 3 puts limits on what an agent can do before anyone has looked.

A PROGRAMME BUILT AROUND PEOPLE

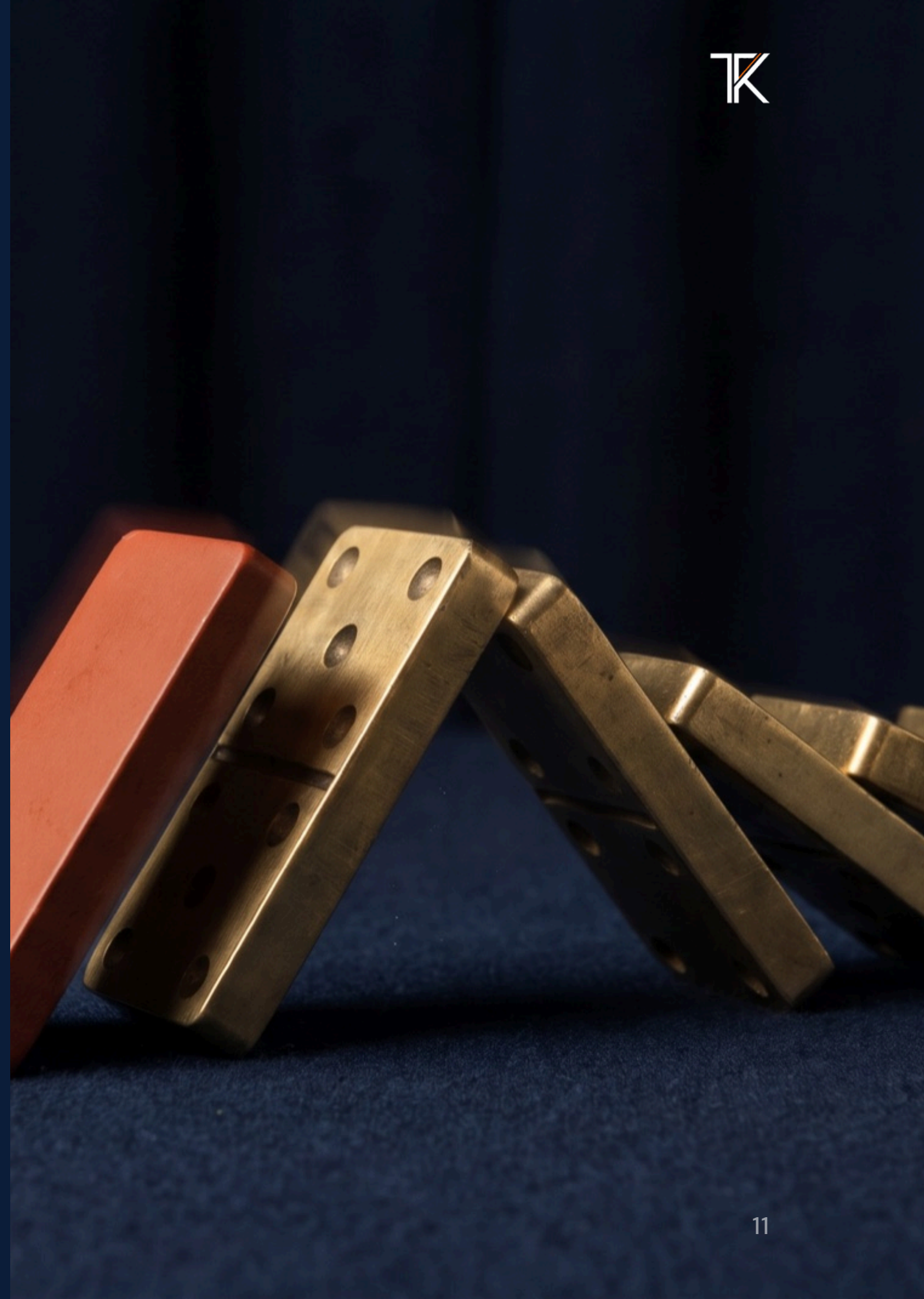
The next morning

When an analyst reviews yesterday's alerts. By then the agent has finished.

02

How damage spreads between agents.

Led by Maria Singh. A compromised laptop is bad. A compromised agent is often worse, for three reasons.



THREE REASONS

Why a compromised agent does more damage.

82:1

Agents run on borrowed authority

Machine identities per human one. 42% of machine identities hold privileged or sensitive access, while 88% of respondents still define a privileged user as a human.

CyberArk survey of 2,600 security leaders, 2025.
Vendor survey.

80%

Agents act beyond their brief

Of companies said their AI agents had taken unintended actions. 23% said agents had been tricked into revealing access credentials.

SailPoint, May 2025. Vendor survey: treat the exact numbers with care.

58–90%

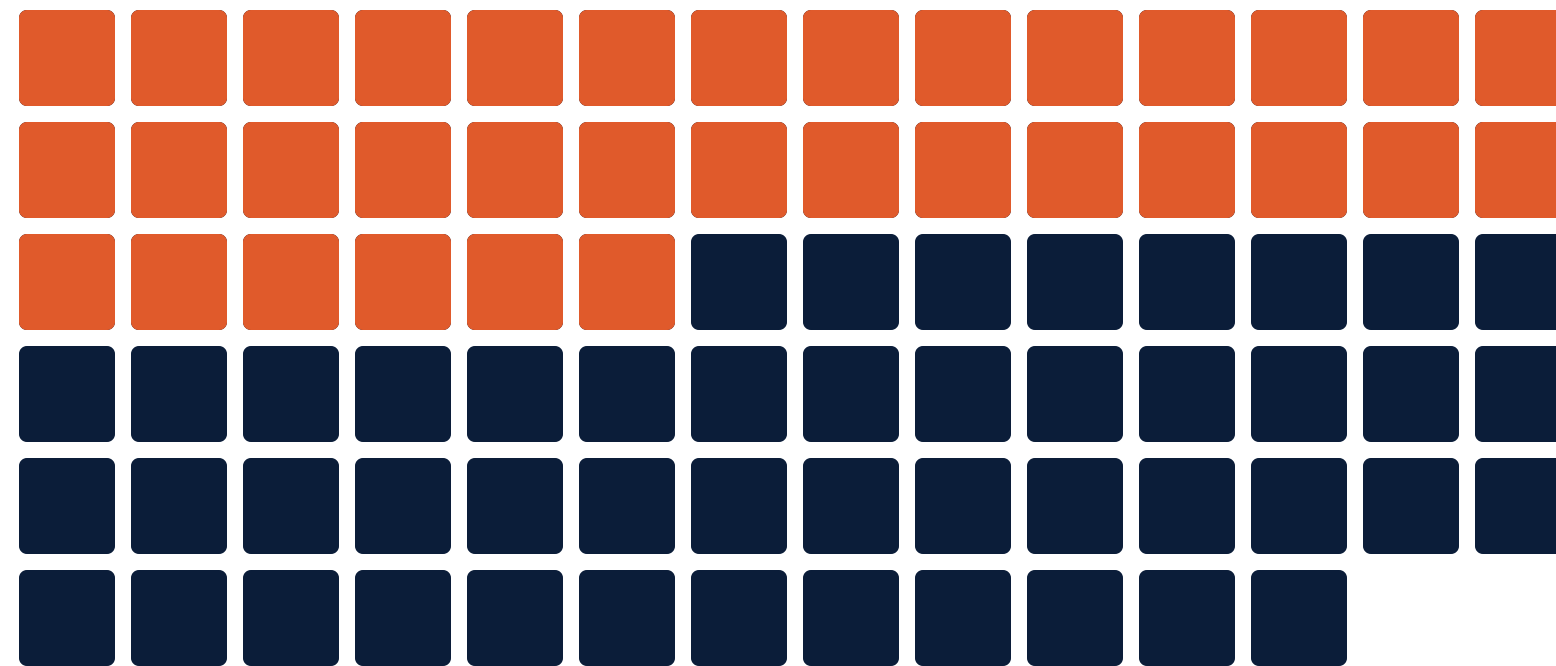
Agents talk to agents

Of trials in which adversarial web content hijacked multi-agent systems into running arbitrary code, even when individual agents refused direct injection.

Cornell Tech: Triedman, Jha, Shmatikov, COLM 2025.
Laboratory.

REASON 1 • BORROWED AUTHORITY

Eighty-two machine identities for every person.



82 machine identities

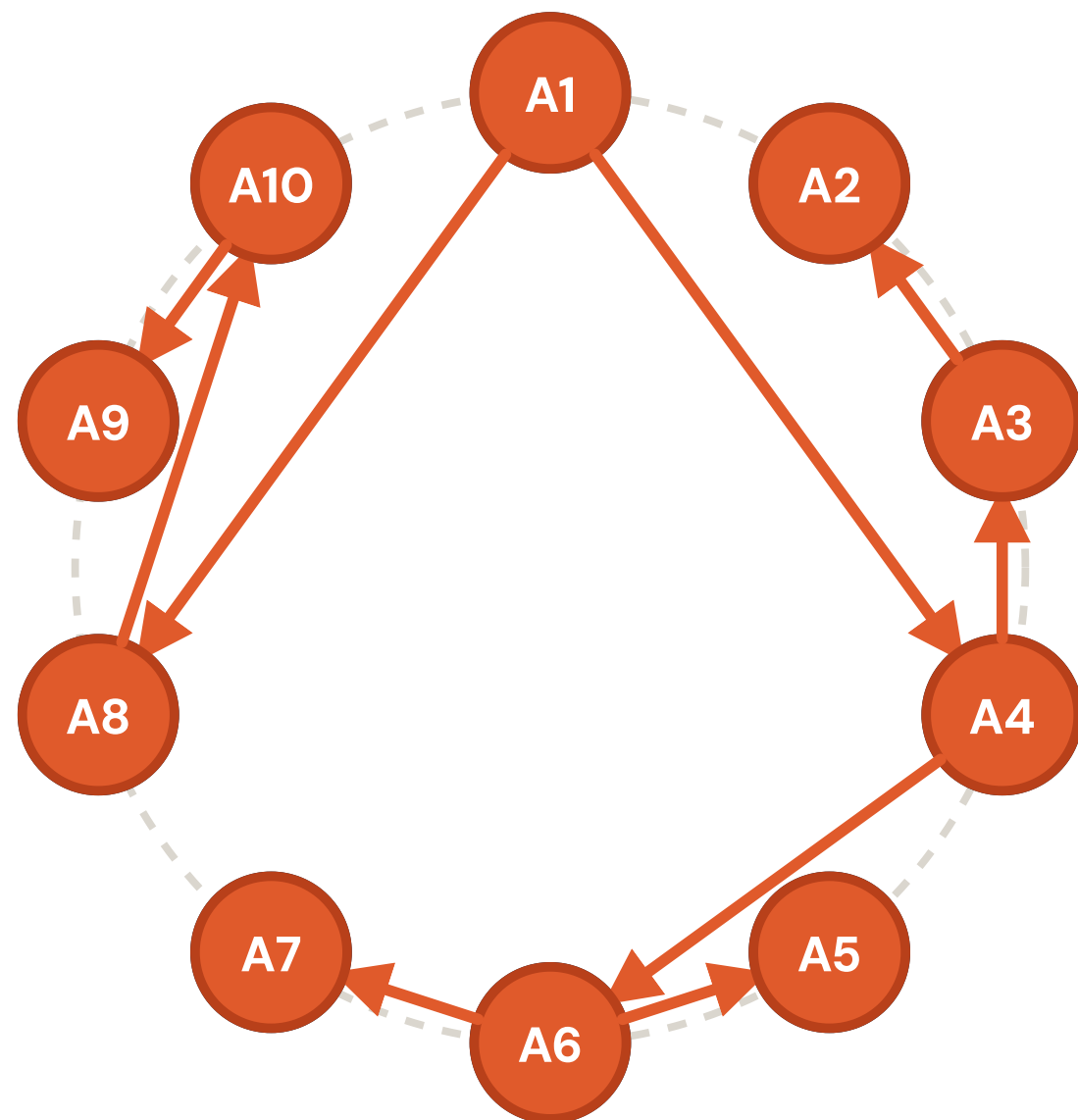
42% with privileged or sensitive access

An agent that inherits a relationship manager's access can do everything that manager can, at machine speed, **and the logs will show the manager's name.**

CyberArk Identity Security Landscape 2025, vendor survey. Grid drawn to scale: 82 cells, 34 of them (about 42%) highlighted.

REASON 3 · AGENTS TALK TO AGENTS

One planted prompt passes from agent to agent.



Turn 5

10 of 10 agents infected

Every agent is infected, by about the fifth turn on average.

Schematic; the order of infection is illustrative. Finding: Lee and Tiwari, Prompt infection, arXiv 2410.07283, a self-replicating prompt infected all ten agents by about turn 4.7. Measured, simulation.

OWASP names these risks directly.

ASIO7

Insecure inter-agent communication

One hijacked agent passes instructions to others that trust it.

ASIO8

Cascading failures

Damage that moves through a multi-agent system turn by turn.

ASI10

Rogue agents

Deleting backups is OWASP's rogue-agent example.

WATCH THE CALL GRAPH

You cannot watch agents one at a time. You need to see which agent invoked which, with whose credentials, carrying what data.

OWASP's mitigations for ASIO8 and ASI10 are the detection controls in this session, plus circuit breakers and kill switches, which Part 3 covers.

03

Twelve action signals to alert on.

Led by Maria Singh. Each compares what an agent is doing against what it normally does or is allowed to do. None depends on reading the agent's words.

ACTION SIGNALS · 1 TO 4

Tempo, new tools and unusual reads.

#	SIGNAL	WHAT IT LOOKS LIKE	WHY IT MATTERS
1	Tool-call tempo	Calls per minute far above the agent's normal rate, or faster than any person could work	The GTG-1002 tell: several operations a second
2	First use of a tool <small>CHEAP TO START</small>	An agent calls a tool, API or connector it has never used, or one outside its declared list	Hijacked agents reach for new capabilities (ASI02)
3	Out-of-scope data read <small>CHEAP TO START</small>	Queries against records, tables or customer segments outside the agent's purpose, or bulk volumes	The usual first step before exfiltration
4	Lots in, little out	Large volumes of data read with little output to the user	Anthropic used this ratio to show the AI was analysing stolen data

Our list. OWASP references are to the Top 10 for Agentic Applications, December 2025.

ACTION SIGNALS · 5 TO 8

Privilege, credentials, instructions and spend.

#	SIGNAL	WHAT IT LOOKS LIKE	WHY IT MATTERS
5	Privilege request	New access scopes, role changes or token exchanges the user did not start	Identity and privilege abuse (ASI03)
6	Credential out of place	An agent's token used from a new workload, address or region, or a person's delegated token used outside their hours	Stolen or misused delegated authority
7	Instruction from content	An action traced to text in an email, page, document or tool output instead of the user's request	The signature of prompt injection (ASI01)
8	Spend velocity	Model spend, API calls or payment value per hour above budget	Cheap, early and hard for an attacker to hide

Our list. OWASP references are to the Top 10 for Agentic Applications, December 2025.

ACTION SIGNALS · 9 TO 12

Fan-out, irreversible verbs, decoys and probing.

#	SIGNAL	WHAT IT LOOKS LIKE	WHY IT MATTERS
9	Agent fan-out	One agent suddenly invokes many others, or a new agent joins the graph	Cascades start here (ASI07, ASI08)
10	Irreversible verbs <small>CHEAP TO START</small>	Deletions, disabled logging, changed guardrails, or edits to the agent's own tools, memory or configuration	The Replit pattern; deleting backups is OWASP's rogue-agent example
11	Decoy touched <small>EARLY RESEARCH</small>	Any read of a canary record, decoy credential or decoy tool	Almost no false positives by design
12	Denials and retries	Repeated authorisation failures followed by a different tool or parameters	An agent probing around its limits

Our list. Signal 11 rests on early research (arXiv 2604.18248, 2606.04141) and vendor tools, not yet on independent production evaluations.

WHERE TO START

Three cheap signals catch a lot.

2 First use of a tool

3 Out-of-scope data read

10 Irreversible verbs

SIGNAL 11 · THE MOST PROMISING NEW IDEA

Decoys turn a subtle problem into an unambiguous alert.

A canary customer record that no legitimate process ever opens. A decoy tool in an agent's tool list that no legitimate task needs.

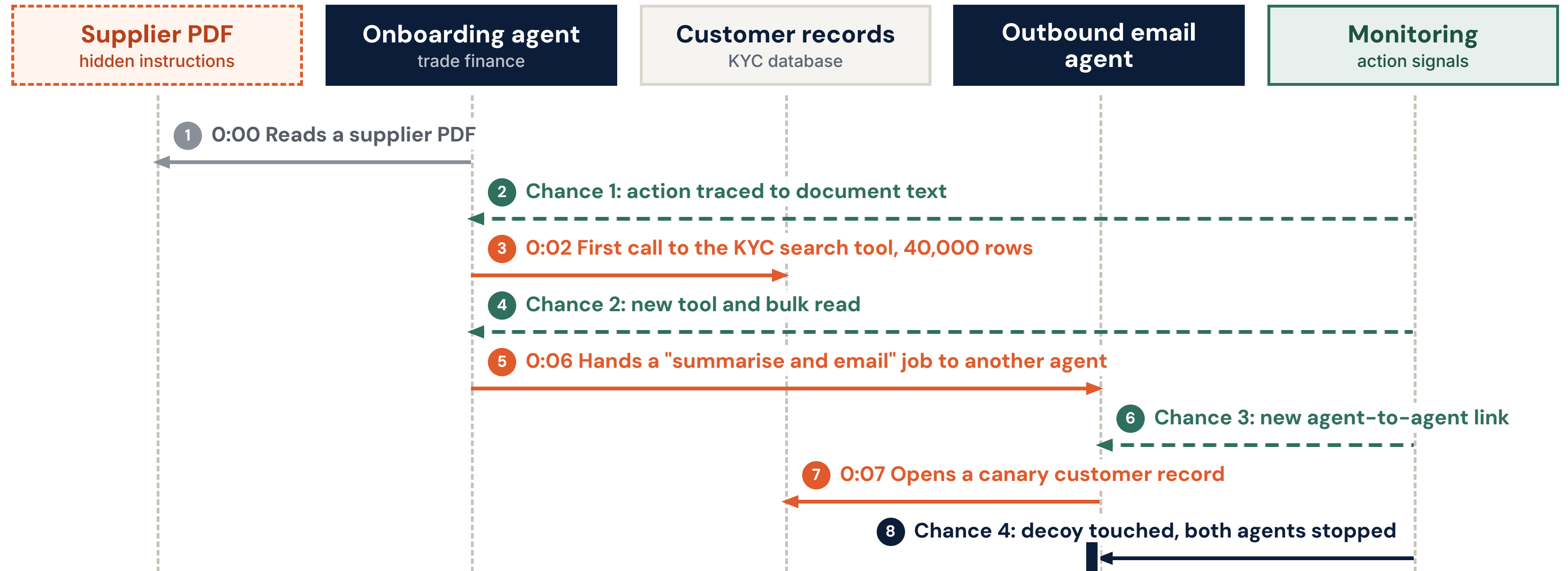
The research on decoys for agents is still early. We would deploy them alongside the other signals rather than instead of them.

04

A rogue agent caught in seven minutes.

A composite walkthrough in a bank, not a real incident. An onboarding agent reads a supplier document with hidden instructions, and monitoring gets its chances to catch what happens next.

Four chances to catch a rogue onboarding agent.

**STEP 8**

Signal 11. A high-confidence alert triggers an automatic stop: both agents' tokens are revoked before the email leaves.

Illustrative composite. Signal numbers refer to the twelve action signals. Orange: the attack. Green: a detection. Navy with a bar: the automatic stop.

WHAT THE WALKTHROUGH SHOWS

Caught at minute seven, before data leaves.

WITH ACTION MONITORING



WITHOUT ACTION MONITORING

**Each agent has its own identity and a declared scope**

So “first use” and “out of scope” mean something.

Someone decided in advance

That a decoy touch stops the agent automatically, without waiting for an analyst. Without that decision, the alert at minute seven is read the next morning.

Composite scenario, illustrative. Dwell time: Mandiant M-Trends 2026. Neither control is exotic.

05

Build the record before you need it.

Led by Terence Kok. None of the signals can fire on data you do not collect.



THE HARDER PART

Nobody can say how many agents they run.

15

Fewer than 15 agents

Run by the average Fortune 500 company in 2025.

150,000

More than 150,000 agents

Forecast for the average Fortune 500 company by 2028.

13%

Think governance is adequate

Of organisations believe their agent governance is adequate.

Whatever the exact numbers turn out to be, an inventory kept by hand will not keep up.

Gartner, 28 April 2026. Grade: forecast. In my experience the detection logic is the easier part; the inventory of agents, identities and access is the harder one.

FOUR PIECES WORTH BUILDING NOW

The record a rogue agent cannot hide from.

1

An identity for every agent

Not a shared service account, and not the identity of the person who launched it.

Microsoft Entra Agent ID (previewed May 2025, generally available through Agent 365 since May 2026); Amazon Bedrock AgentCore Identity (generally available October 2025); per-agent identities in Gemini Enterprise.

2

A log of every action, keyed to that identity

OpenTelemetry's generative AI conventions define spans for invoking an agent and executing a tool.

Still marked as in development, so expect changes. MITRE ATLAS lists AI telemetry logging (AML.M0024) among its mitigations.

3

A declared scope for every agent

The tools, data and actions it is meant to use.

Without one, signals 2, 3 and 5 have nothing to compare against.

4

Provenance for instructions

Whether each action followed the user's request or something the agent read.

The hardest of the four to build, and the one that catches prompt injection directly.

AFTER ANY INCIDENT

“What did the agent do, on whose authority, and when did you know?”

The first question any regulator or board will ask. The four records answer it.

06

The clocks start at discovery.

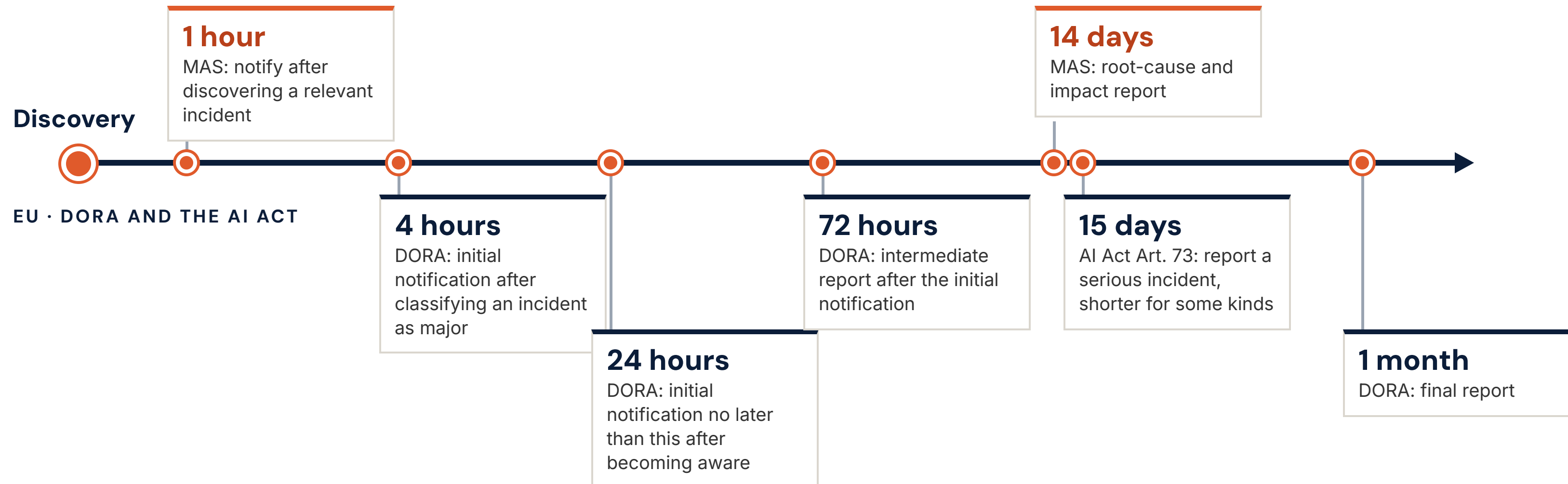
Led by Terence Kok. For a financial institution, detecting a rogue agent starts a regulatory clock.



SINGAPORE AND THE EU

The reporting clocks after discovery.

SINGAPORE · MAS

**AI ACT ART. 26**

Deployers of high-risk systems monitor their operation, suspend use and inform the provider if they suspect a risk, and keep the automatically generated logs for at least six months. **6 months →**

MAS Notice FSM-N07; Delegated Regulation (EU) 2025/301 (DORA); Regulation (EU) 2024/1689. Spacing is schematic. DORA's 72-hour and one-month clocks run from earlier reports.

WHAT SUPERVISORS NOW EXPECT

Supervisors expect monitoring and oversight.

MAS, 7 Oct 2026	Guidelines on AI Risk Management cover AI agents and expect monitoring and human oversight across the lifecycle.	MAS
CSA, June 2026	Singapore's addendum on securing agentic AI lists rogue actions among the risks it addresses.	CSA
Five Eyes, 2026	Joint guidance on agentic AI services: give agents distinct identities, log every action, assume agents may behave unexpectedly.	JOINT
EU AI Act Art. 26	Deployers monitor high-risk systems, suspend use and inform the provider if they suspect a risk, and keep logs for at least six months.	EU
EU AI Act Art. 73	A 15-day deadline for reporting serious incidents, shorter for some kinds.	EU

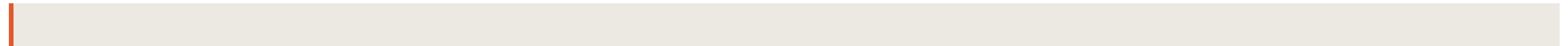
PUT THE CLOCKS NEXT TO THE DWELL TIMES

The record makes the one-hour notification possible.

14 days Median dwell before an intruder is detected



1 hour MAS notification after discovery, drawn to the same scale



An institution that cannot reconstruct what an agent did cannot say within an hour whether an incident is reportable, **let alone describe it.**

Different clocks: dwell time runs before discovery, the MAS clock after it. Mandiant M-Trends 2026; MAS Notice FSM-N07. One hour is 1/336 of 14 days.

WHO ACTS WHEN THE ALARM GOES

Three decisions to make in advance.

DECISION 1

Who may stop an agent

Name the role with authority to suspend an agent in production, and make sure it is someone on shift at three in the morning. For the highest-risk agents, let specific signals stop the agent automatically.

DECISION 2

What stopping means

Revoke the agent's tokens, halt its queue, cut its tool access and freeze anything it handed to other agents. A stop that leaves delegated credentials alive is not a stop.

DECISION 3

What to preserve

The action log, the agent's memory and context, its tool definitions at the time, and the content it read: the evidence for the root-cause report MAS expects within 14 days.

REHEARSE IT

A tabletop exercise built on the walkthrough takes an afternoon and will show whether the one-hour notification is realistic.

07

For the board.

Five questions for management, four for this room, and how strong the evidence is.

QUESTIONS FOR THE BOARD

Ask management for written answers.

- 1** How many AI agents do we run in production, and does each have its own identity?
 - 2** Do we log what our agents do, including tool calls, data read and credentials used, or only what they say?
 - 3** Which of the twelve signals can our security operations centre alert on today?
 - 4** Who can stop an agent in production at any hour, and when did we last test it?
 - 5** If an agent went rogue this morning, could we tell MAS within an hour what it did?
-

The fifth question is the test of the other four.

FOR DISCUSSION

Four questions for this room.

- 1 Which of our agents read outside content, such as emails, documents or web pages, and can act on it?
- 2 If an agent misbehaved at three in the morning, who would be called, and could they stop it?
- 3 Which of our agents act on a person's credentials, so the logs would show that person's name?
- 4 Where would a decoy record or a decoy tool give us an alert we would trust at once?

How strong each claim is.

Claim	Source	Grade
GTG-1002 detected because "the sustained nature of the attack triggered detection"; tempo and volume analysis	Anthropic report, November 2025	Reported, first-party
Median dwell time 14 days; access-to-hand-off 22 seconds (from over 8 hours in 2022)	Mandiant M-Trends 2026	Measured
Average breakout 29 minutes; fastest 27 seconds	CrowdStrike Global Threat Report 2026	Measured, vendor
Breach lifecycle 241 days; 97% of organisations with AI breaches lacked AI access controls	IBM Cost of a Data Breach 2025	Measured, survey
82 machine identities per human; 42% privileged; 88% define privileged users as humans	CyberArk Identity Security Landscape 2025	Vendor survey
80% report unintended agent actions; 23% saw agents tricked into revealing credentials	SailPoint, May 2025	Vendor survey
Multi-agent hijacking via web content in 58-90% of trials	Triedman, Jha, Shmatikov, COLM 2025	Measured, laboratory
Self-replicating prompt infected all ten agents by about turn 4.7	Lee and Tiwari, arXiv 2410.07283	Measured, simulation

Forecasts, frameworks and legal texts.

Claim	Source	Grade
Agents per Fortune 500 firm: under 15 (2025) to over 150,000 (2028); 13% think governance adequate	Gartner, 28 April 2026	Forecast
OWASP ASI07, ASI08, ASI10	OWASP Top 10 for Agentic Applications, December 2025	Framework
OpenTelemetry agent and tool spans, status in development	OpenTelemetry semantic conventions for generative AI	Standard, draft
Entra Agent ID, Bedrock AgentCore, Gemini Enterprise agent identity	Vendor announcements, 2025-2026	Reported
MAS one-hour notification, 14-day root-cause report	MAS Notice FSM-N07	Regulatory text
DORA: 4 hours from classification, 24 hours from awareness, 72 hours, one month	Delegated Regulation (EU) 2025/301	Regulatory text
AI Act Art. 26 monitoring and six-month logs; Art. 73 fifteen days	Regulation (EU) 2024/1689	Legal text
Decoys and canaries for agents	arXiv 2604.18248, 2606.04141; vendor tools	Early research

EVIDENCE AND METHODOLOGY

Where we could be wrong.

Speed and dwell figures come from incident-response firms' own case data and measure different things, so we show them side by side and never combine them. Identity and agent-behaviour figures are vendor surveys. Multi-agent findings are laboratory results. The twelve signals, the walkthrough and the questions are ours; the walkthrough is illustrative.

IF PLATFORMS CHANGE

If agent platforms build action monitoring and automatic stops into their defaults, much of what we describe as a project becomes a setting, and the hard part shifts to deciding thresholds.

IF THE SURVEYS OVERSTATE

If independent data shows unintended actions are rarer than vendor surveys suggest, the case for automatic stops on lower-risk agents weakens. It would not weaken for agents that can move money or release personal data.

SOURCES · 1 OF 2

Numbered as in the post.

- 1 Anthropic. (2025, November 13). Disrupting the first reported AI-orchestrated cyber espionage campaign, and full report (PDF).
anthropic.com
- 2 Google Cloud (Mandiant). (2026, March 24). M-Trends 2026.
cloud.google.com
- 3 CrowdStrike. (2026, February 24). 2026 CrowdStrike Global Threat Report.
crowdstrike.com
- 4 IBM. (2025, July 30). IBM report: 13% of organizations reported breaches of AI models or applications, 97% of which reported lacking proper AI access controls.
newsroom.ibm.com
- 5 CyberArk. (2025, April 23). 2025 Identity Security Landscape (Business Wire).
businesswire.com
- 6 SailPoint. (2025, May). AI agents: the new attack surface.
sailpoint.com
- 7 Triedman, H., Jha, R., & Shmatikov, V. (2025). Multi-agent systems execute arbitrary malicious code. arXiv 2503.12188; COLM 2025.
arxiv.org
- 8 Lee, D., & Tiwari, M. (2024). Prompt infection: LLM-to-LLM prompt injection within multi-agent systems. arXiv 2410.07283.
arxiv.org
- 9 Gartner. (2026, April 28). Gartner identifies six steps to manage AI agent sprawl.
gartner.com
- 10 OWASP GenAI Security Project. (2025, December 9). OWASP Top 10 for Agentic Applications for 2026.
genai.owasp.org

SOURCES · 2 OF 2

Platforms, regulators and guidance.

- | | |
|--|--|
| <p>11 MITRE. ATLAS data releases (v2026.09).
github.com</p> <hr/> <p>12 OpenTelemetry. Semantic conventions for generative AI agent spans.
github.com</p> <hr/> <p>13 AWS. (2025, October 13). Amazon Bedrock AgentCore is now generally available.
aws.amazon.com</p> <hr/> <p>14 Microsoft. Microsoft 365 roadmap: Agent 365.
microsoft.com</p> <hr/> <p>15 Monetary Authority of Singapore. Notice FSM-N07 on technology risk management, and Circular on financial institution incident reporting (December 2025).
mas.gov.sg</p> <hr/> | <p>16 Monetary Authority of Singapore. (2026, October 7). MAS sets out supervisory expectations on responsible AI adoption by financial institutions.
mas.gov.sg</p> <hr/> <p>17 CMS. (2025, October). Cyber incident response and compliance with DORA.
cms.law</p> <hr/> <p>18 EU AI Act, Article 26: Obligations of deployers of high-risk AI systems.
artificialintelligenceact.eu</p> <hr/> <p>19 Cyber Security Agency of Singapore. (2026). Securing agentic AI: addendum to the Guidelines on Securing AI Systems.
csa.gov.sg</p> <hr/> <p>20 Crowell & Moring. (2026). American and allied cyber agencies issue first joint guidance on securing agentic AI.
crowell.com</p> <hr/> |
|--|--|

THANK YOU

Follow the actions.

Read the full series at
terencekok.com/blog/series/rogue-ai-agents/

Terence Kok and Maria Singh

Part 3 turns to the controls: autonomy tiers, hard limits and kill switches.



Read this post