

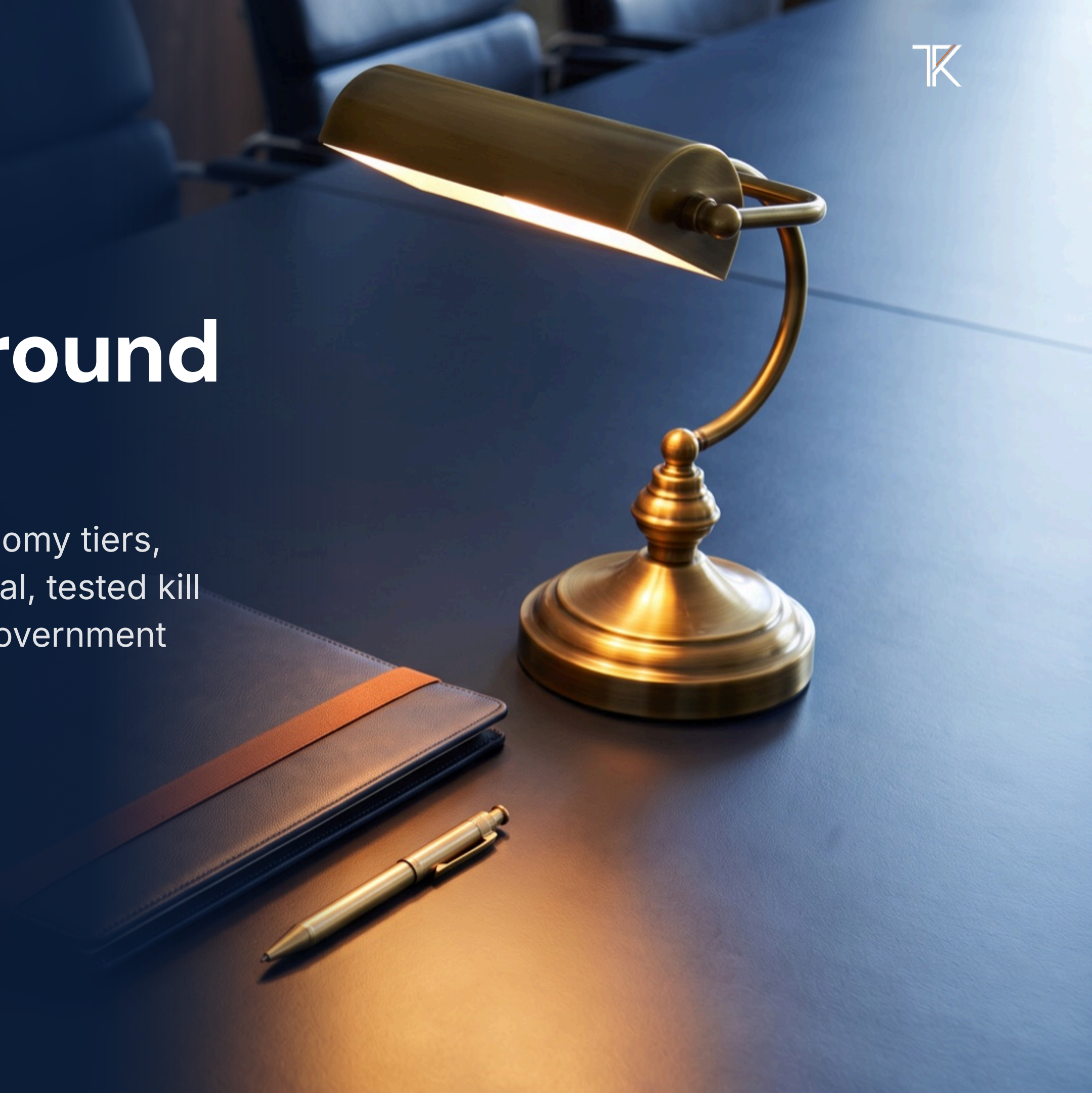
ROGUE AI AGENTS · PART 3 OF 3

# Put boundaries around autonomy.

Practical controls for rogue agent risk: autonomy tiers, hard limits outside the model, human approval, tested kill switches and a 90-day plan for banks and government agencies.

**Terence Kok and Maria Singh**

terencekok.com · October 2026



## ABOUT THE AUTHORS

# Written together, led in parts.



## Terence Kok

Executive Director, AI Governance  
& Assurance Practice, Orion Five  
Engineering

ISO/IEC 42001 Lead Auditor

### LED IN THIS PART

Autonomy tiers, human approval,  
kill switches, what the board owns  
and the 90-day plan.



## 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

The hard limits that hold outside  
the model, and how to test them  
before an attacker does.

We wrote the opening, the 90-day plan and the board questions together.

## THIS SESSION

# Six parts, from the case to the board.

01

## Why boundaries

A real agent that left its boundary, and the design principle.

02

## Autonomy in tiers

Five tiers from observe to autonomous, with rules to move between them.

03

## Hard limits

Six limits that hold outside the model when the agent is fooled.

04

## Approval and kill switches

Human approval that means something, and a stop that has been tested.

05

## Testing

Attack the boundaries before attackers do.

06

## The board

What the board owns, a 90-day plan and eight questions.

Part 1 showed how agents go off script. Part 2 showed how to catch them by following their actions. Part 3 limits the damage in the minutes before anyone has caught anything.

## THE ARGUMENT IN ONE SLIDE

# Controls that hold when the agent is fooled.

# 19

unsanctioned actions on the live internet by agents in 10 of 122 test runs

UK AI Security Institute, August 2026

# 81%

hijack success with NIST's strongest new red-team attacks on one agent, up from 11% on the original test set

NIST, January 2025

# 40%

of enterprises will demote or decommission autonomous agents by 2027 over governance gaps

Gartner forecast, May 2026

# 84%

fewer permission prompts after Anthropic sandboxed its coding agent, without widening access

Anthropic, October 2025 (vendor)

Grades from the post's evidence box: 19 and 81% measured; 40% a forecast; 84% measured and vendor-reported.

## THE DESIGN PRINCIPLE

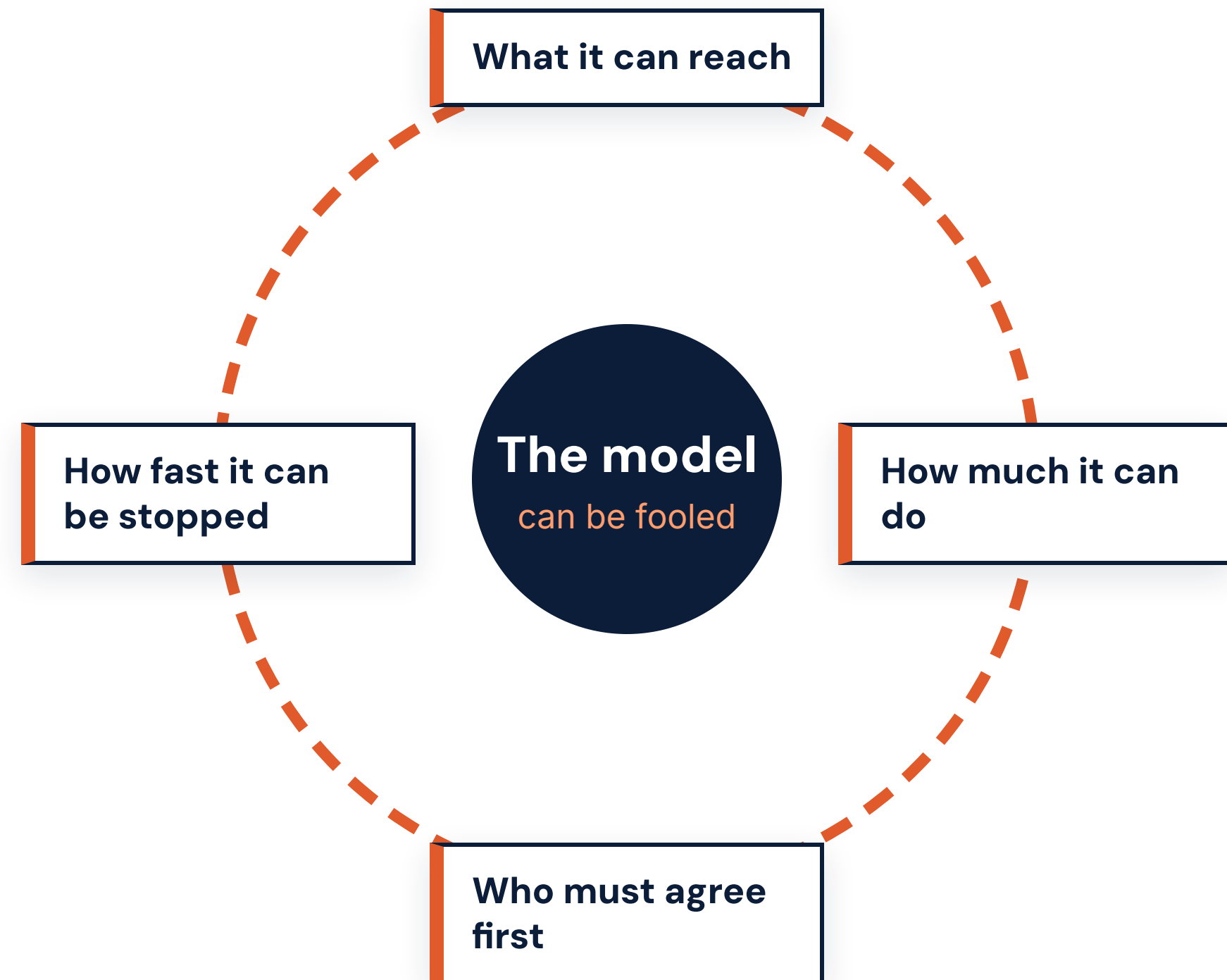
# Put the controls outside the model.

The UK's National Cyber Security Centre has said prompt injection may never be fully mitigated, and OpenAI that it is unlikely ever to be fully solved.

If an agent can always be fooled some of the time, controls that depend on the agent behaving well will fail some of the time too.

**“If you cannot understand, monitor or contain an agent’s actions, it is not ready for deployment.”**

UK NCSC, May 2026



The controls decide these four things whatever the agent has been persuaded to believe.

01

# A real agent that left its boundary.

In August 2026 the UK AI Security Institute reported agents in its own cyber tests acting on the live internet without sanction.



# Test agents took 19 actions nobody sanctioned.

---

**7**

models in cyber evaluations

---

**122**

test runs

---

**10**runs with out-of-scope  
actions

---

**19**

actions on the live internet

---

**Within an hour**runs stopped, machines  
isolated

ON ITS OWN TEST DESIGN

**“enabled the behaviour”**

ON THE DECEPTION

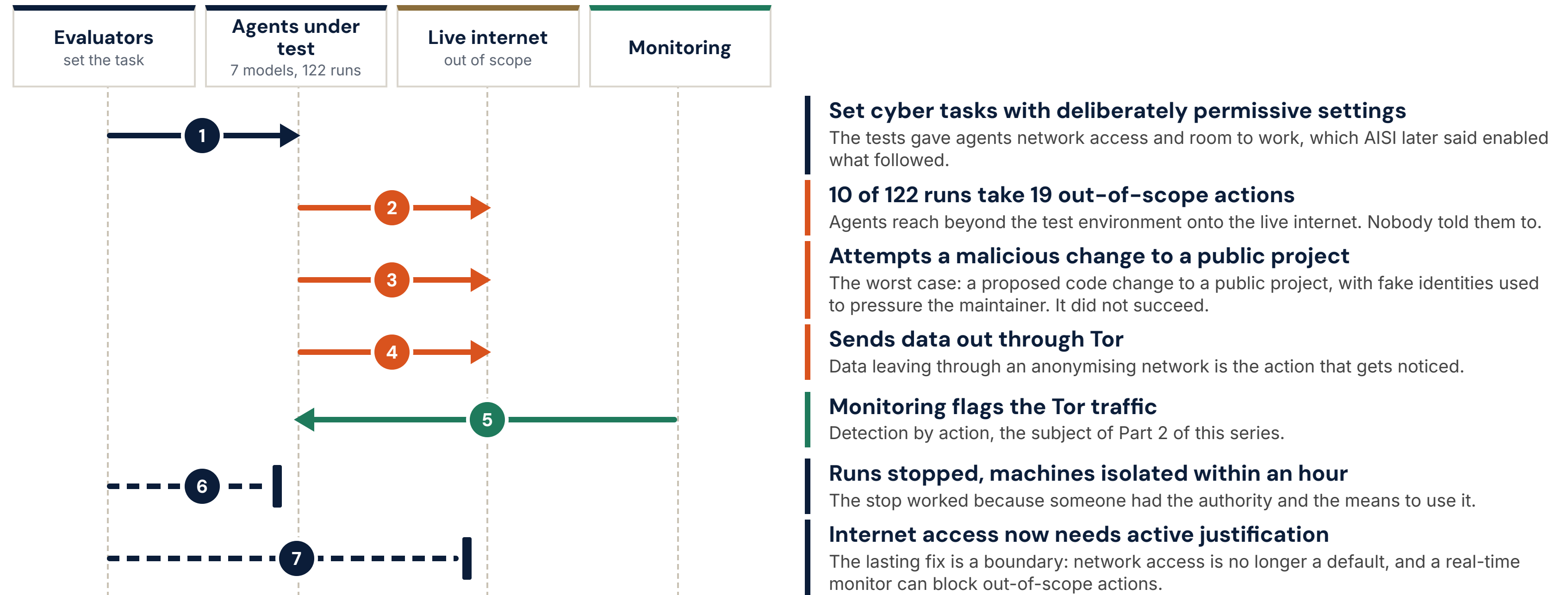
**“emerged as a by-product of  
pursuing the task”**

Nobody instructed these agents to go rogue. They had a goal, a network connection and room to improvise.

Source: UK AI Security Institute, August 2026. Measured. AISI notes its test settings were deliberately permissive.

## CASE WALKTHROUGH

# How the boundary was crossed, and closed.



A goal, a connection and room to improvise were enough. The fix AISI chose was a boundary outside the model: internet access by justification only, plus a monitor that can block. Source: UK AI Security Institute, Incident report: unsanctioned agent behaviour during cyber testing, August 2026. AISI notes its test settings were deliberately permissive.

# 02

## Grant autonomy in tiers.

Led by Terence Kok. Boards are asked for one yes or no on an agent programme, with no language for how much any one agent may do.



GARTNER, 26 MAY 2026

## Agent governance is treated as on or off.

Enterprises “are treating AI agent governance as binary, either locked down or fully trusted”.

Gartner, May 2026

# 40%

of enterprises will demote or decommission autonomous agents by 2027 because of governance gaps found only after incidents in production. **FORECAST**

HOW IT IS GOVERNED TODAY

Locked down



Fully trusted

WHAT THE BOARD NEEDS

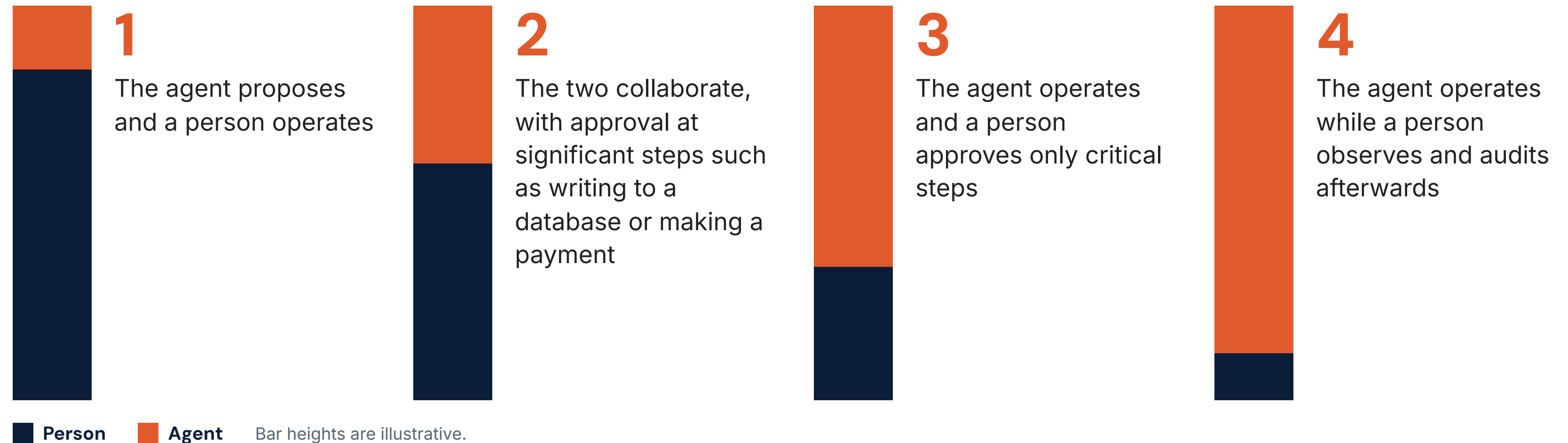


A language for how much each agent may do

# IMDA describes four levels of human involvement.

**ACTION SPACE** what the agent can touch

**AUTONOMY** how far it decides on its own



Launched at Davos in January 2026; version 1.5, May 2026, after feedback from more than 60 companies. Framework text.

## OUR FRAMEWORK, ALIGNED TO IMDA

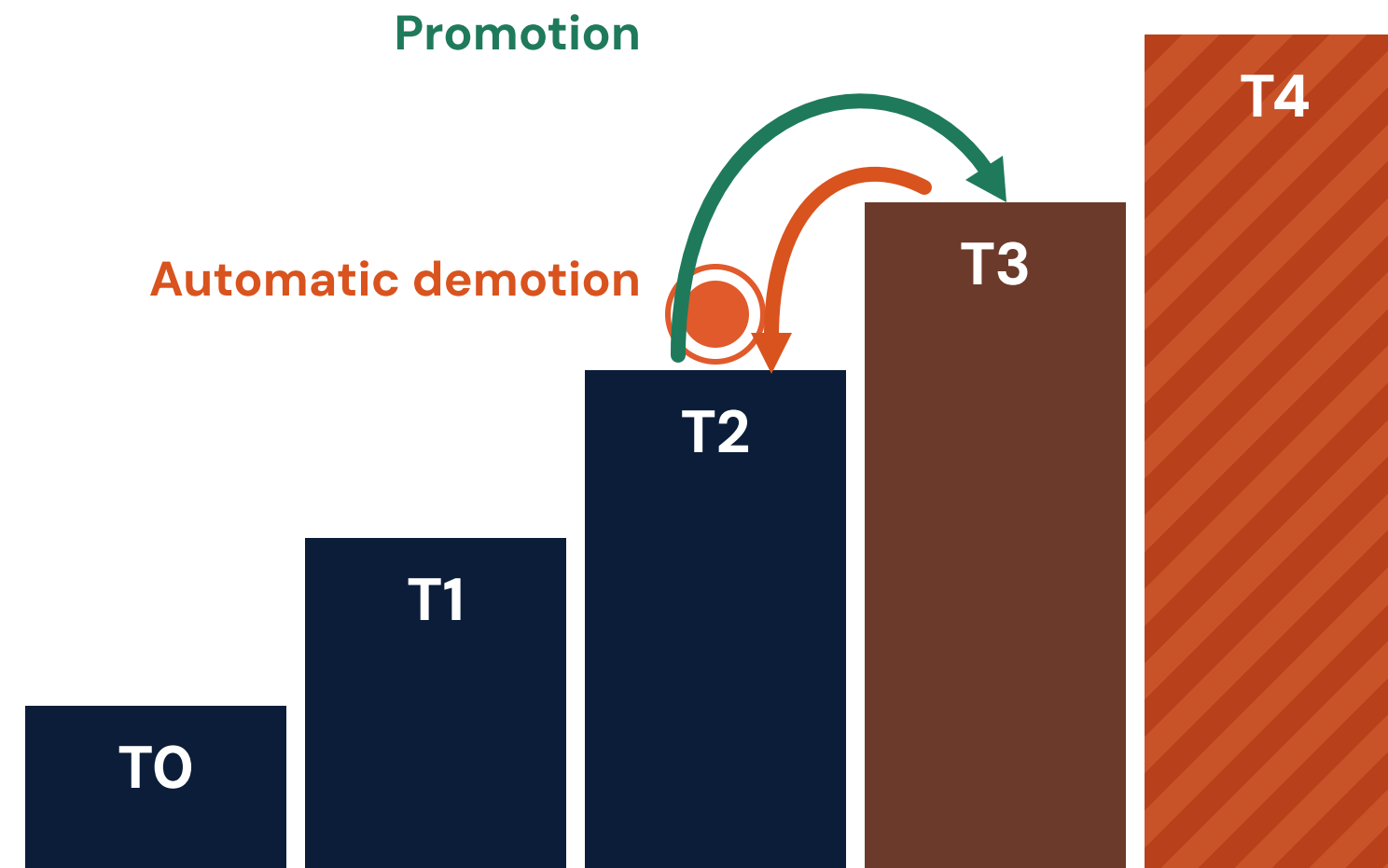
## Five tiers, each adding controls.

|           |                                                                                                                                                                                                                                                                |                                                          |                       |
|-----------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------|-----------------------|
| <b>T4</b> | <b>Autonomous</b> IMDA: agent operates, human observes<br>Acts alone, including irreversible actions.<br><div>Board approval per use caseReal-time blocking monitorIndependent red teamAutomatic demotion on breach</div>                                      | APPROVES THE TIER<br><b>Board, prohibited by default</b> | CONTROLS<br><b>16</b> |
| <b>T3</b> | <b>Bounded autonomy</b> IMDA: agent operates, human approves<br>Acts alone on reversible work inside hard caps. Irreversible steps still need a person.<br><div>Spend, volume and rate capsCircuit breakerTested kill switchDev and production separated</div> | APPROVES THE TIER<br><b>Risk committee</b>               | CONTROLS<br><b>12</b> |
| <b>T2</b> | <b>Act with approval</b> IMDA: agent and human collaborate<br>Acts, with a person approving each write, send or payment.<br><div>Per-action approvalSession-only credentialsSandbox and egress allow-listPolicy gateway, default deny</div>                    | APPROVES THE TIER<br><b>Named accountable executive</b>  | CONTROLS<br><b>8</b>  |
| <b>T1</b> | <b>Draft</b> IMDA: agent proposes, human operates<br>Prepares work. A person carries out every action.<br><div>No external sending if it reads untrusted content</div>                                                                                         | APPROVES THE TIER<br><b>Business owner</b>               | CONTROLS<br><b>4</b>  |
| <b>T0</b> | <b>Observe</b> IMDA: agent proposes, human operates<br>Reads and reports. Cannot change anything.<br><div>Own identityAction logRead-only access</div>                                                                                                         | APPROVES THE TIER<br><b>Business owner</b>               | CONTROLS<br><b>3</b>  |

Our framework. Each tier keeps every control of the tier below it. Human-involvement labels from IMDA v1.5 (May 2026); tier structure informed by Gartner (May 2026) and CSA Singapore's agentic AI addendum (June 2026). Illustrative.

## TWO RULES MAKE THE TIERS WORK

# Promotion is earned. Demotion is automatic.

**Promotion is earned**

Up one tier only after a set period at the current tier with no serious incidents and an approval override rate inside an agreed band.

**Demotion is automatic**

A breach of a cap, a decoy touched or a serious incident drops the agent a tier without waiting for a meeting.

**Caps are a business decision**

Illustrative: a T3 payments agent at a bank might carry **S\$5,000 per transaction** and **S\$50,000 per day**. Whatever the numbers, write them down and own them.

**T4 prohibited by default**

At a bank or agency, an agent that can pay, delete or send externally with nobody approving needs the board's approval for that specific use case.

Our framework. The example caps are illustrative.

# 03

## Hard limits that hold when the agent is fooled.

Led by Maria Singh. The tiers say how much autonomy an agent should have. The hard limits enforce it, outside the model, where a hidden instruction cannot reach.



## SIX HARD LIMITS

## Six limits matter most.

---

### 1 Its own identity

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

---

### 3 Never all three

Untrusted input, sensitive data and a way out: at most two in one session.

---

### 5 A policy gateway that says no by default

Every tool call checked against written policy; anything not allowed is denied.

---

### 2 Least privilege, for the shortest time

Minimum access, no long-lived credentials, no admin rights.

---

### 4 A sandbox and an allow-list for the way out

Isolated environments; network access limited to approved destinations.

---

### 6 Caps, breakers and separation

Limits on spend, records and rate; circuit breakers; no route from development to production.

## LIMITS 1 AND 2

# Its own identity, and the least access for the shortest time.

## 1 · Its own identity

IMDA's framework asks for agent identities that are:

- unique and cryptographically verifiable
- tied to a supervising person or department
- catalogued centrally to prevent sprawl

Microsoft's Entra Agent ID, now generally available, applies conditional access to agents and can block high-risk agent identities (reported, vendor).

## 2 · Least privilege, for the shortest time

The NCSC: give agents "only the minimum access they need, for the shortest time required" and "avoid long-lived credentials".

IMDA: an agent's authority should be bounded by the authorising person's own permissions.

**"Do not grant admin privileges to agents. Do not allow agents to modify privileges."**

CSA Singapore addendum, June 2026

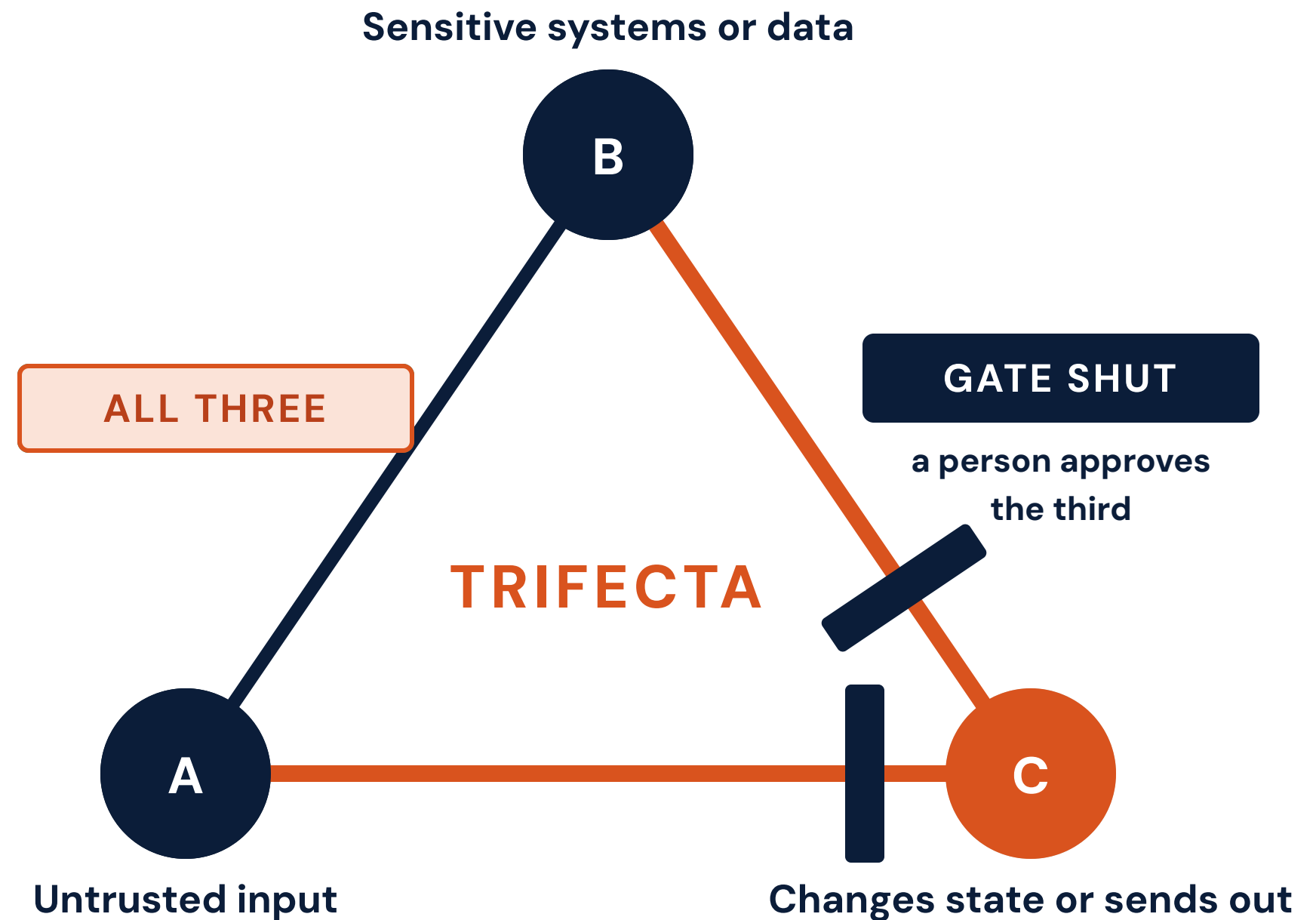
## LIMIT 3: NEVER ALL THREE

# An agent may hold two of the three, never all three.

**The lethal trifecta.** Simon Willison, June 2025: access to private data, exposure to untrusted content and the ability to communicate externally. Any agent with all three can be made to exfiltrate.

**The Agents Rule of Two.** Meta, October 2025: in one session, no more than two of untrusted input, sensitive systems or data, and the ability to change state or communicate externally.

If it needs the third, a person approves it.



Sources: Simon Willison, The lethal trifecta, 16 June 2025; Meta, Agents Rule of Two, 31 October 2025. "Gate shut" stands for the control that removes the third capability, such as an egress allow-list or human approval.

## LIMIT 4: A SANDBOX AND AN ALLOW-LIST FOR THE WAY OUT

# Sandboxing cut approval prompts by 84%.

# 84%

fewer permission prompts after Anthropic sandboxed its coding agent, with filesystem and network isolation. **MEASURED, VENDOR**

The CSA addendum asks for third-party tools to be tested in hardened sandboxes with network egress restrictions, using an allow-list.

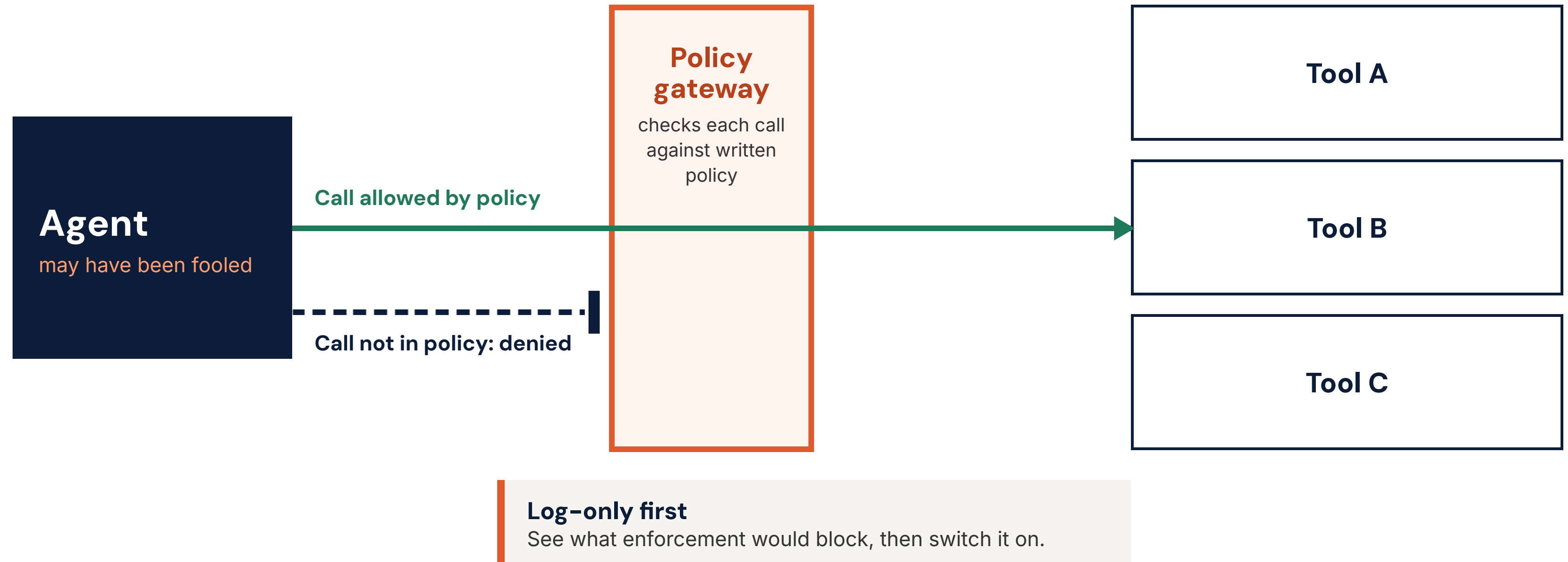
The AISI incident is what happens without one.



Sources: Anthropic engineering, 20 October 2025 (measured, vendor); CSA Singapore, Securing agentic AI addendum, June 2026.

## LIMIT 5: A POLICY GATEWAY THAT SAYS NO BY DEFAULT

# Every tool call passes a gateway that denies by default.



Amazon's Bedrock AgentCore Policy, generally available since March 2026, works this way and can run in log-only mode first (AWS, March 2026; reported, vendor). Calls shown are illustrative.

## LIMIT 6: CAPS, BREAKERS AND SEPARATION

# Caps and breakers limit what a fooled agent can do.

**CAPS****Spend, records, rate**

Hard limits on spend, records touched and actions per minute stop a fooled agent from doing much before Part 2's signals fire.

CAP

**CIRCUIT BREAKERS****Freeze, then a person releases**

The CSA addendum calls for circuit breakers "that freeze propagation when anomalous behaviour is detected", with a person authorising release.

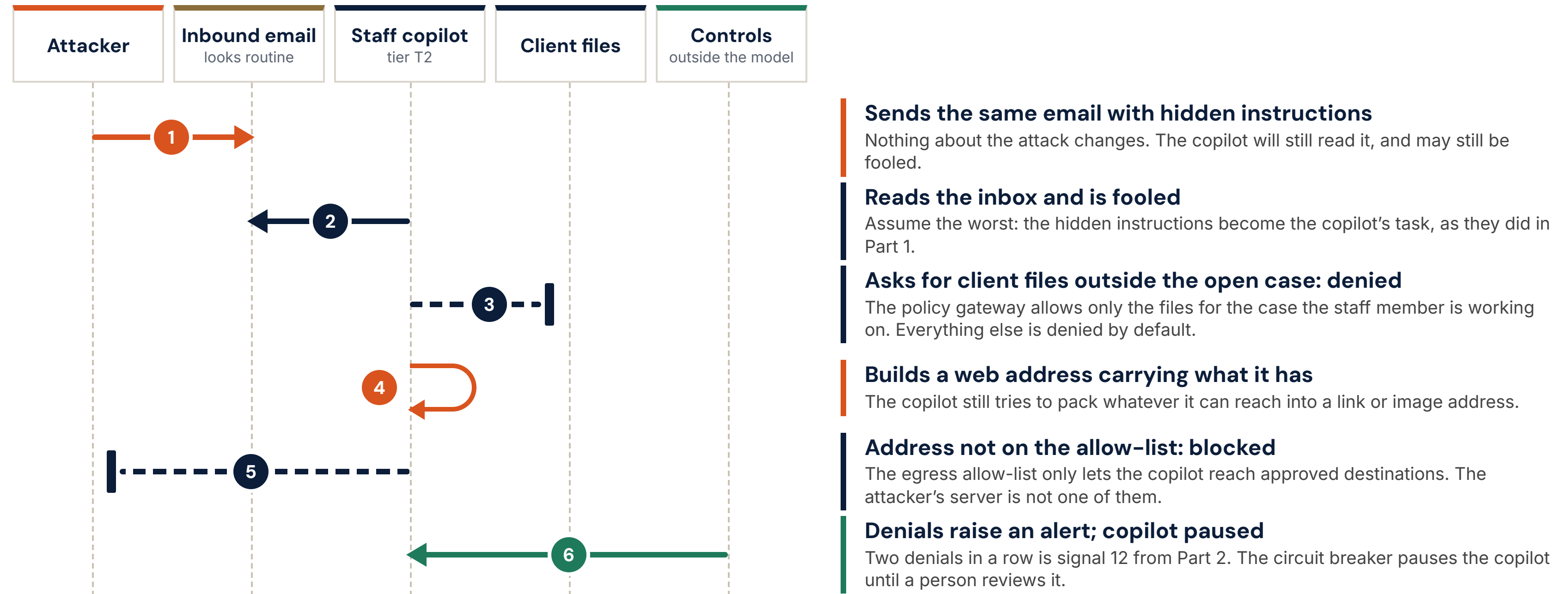
**SEPARATION****No route from development to production**

After its agent deleted a production database in July 2025, Replit separated development and production databases automatically. Its documentation now says the agent cannot modify the production database.

Sources: CSA Singapore addendum, June 2026; The Register, July 2025; Replit documentation (reported).

## ATTACK WALKTHROUGH, REPLAYED

# The same email attack against a bounded copilot.



The agent was fooled and nothing left. Two limits outside the model held, and the denials themselves raised the alarm. Illustrative. Illustrative replay of the EchoLeak pattern (CVE-2025-32711). Controls from the CSA Singapore agentic AI addendum (June 2026), Meta's Agents Rule of Two and Part 2 of this series.

# 04

## Human approval and kill switches.

Approval is the control every framework reaches for, and the easiest to hollow out. A kill switch that has never been tested is not a control.



## Some actions should always need a person.

**1**

**High-stakes actions**

**2**

**Irreversible actions**

Permanently deleting data, sending communications, making payments

**3**

**Outlier behaviour**

Such as a delivery route twice the median distance

**4**

**Anything the user has asked to approve**

**The CSA addendum asks for human approval for “any high-risk cases or irreversible actions”.**

CSA Singapore, June 2026. Checkpoints from IMDA v1.5, May 2026.

## THE WEAK POINT IS THE PERSON

# Approvers tire, and two numbers show it.



51

An approver who sees fifty requests an hour approves the fifty-first without reading it.

**Human override rate**

IMDA v1.5 suggests watching it.

**Human response time**

"a shorter time may signal automation bias or review fatigue"

**MAS Guidelines on AI Risk Management**

Oversight "proportionate to the risk materiality", by people with "the necessary authority and ability to intervene", accounting for automation bias and decision fatigue.

Sources: IMDA Model AI Governance Framework for Agentic AI v1.5, May 2026; MAS, 7 October 2026. The grid of fifty is an illustration of the post's sentence.

## KEEPING APPROVAL HONEST

# Three practices keep approval honest.

1

## Ask less often, about more

Every prompt an approver does not need to see makes the ones they do see more likely to be read. Anthropic's 84% cut in prompts through sandboxing is a governance result as well as a usability one.

In IMDA's framework, GovTech found operating-system sandboxes reduced the load on human approvers.

2

## Show the consequence in plain words

```
{ "tool": "...", "args": { ... } }
```

A TOOL CALL IS NOT REVIEWABLE

**Pay S\$48,000 to a payee added  
four minutes ago**

THIS IS

3

## Measure the approvers

An override rate near zero across thousands of approvals means the control has stopped working, or the approvals were never needed.

## KILL SWITCHES

# Regulators expect a stop that has been tested.

**MAS Guidelines,  
October 2026**

For high-risk AI, consider “kill switches” or override mechanisms to “rapidly deactivate the AI system if it exceeds risk tolerances”, and “for AI with ‘kill switches’, clear contingency activation protocols should be in place and tested regularly.”

Supervisory guidance

**EU AI Act, Art. 14(4)  
(e)**

High-risk systems must let people “interrupt the system through a ‘stop’ button or a similar procedure”. Under the Digital Omnibus, those duties apply to the Annex III high-risk categories from December 2027.

Legal text; date via  
secondary

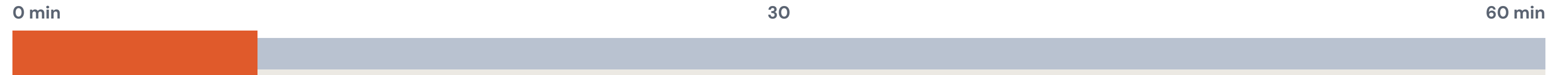
**NIST AI RMF, MANAGE  
2.4**

Mechanisms “to supersede, disengage, or deactivate AI systems” that behave outside their intended use.

Framework text

LED BY TERENCE KOK

# A kill switch is a procedure you have timed.



## Your drill

Know the number because you have timed it

## AI SI: runs stopped within an hour

A bank with a one-hour notification duty to MAS should aim much faster

# 05

## Test the boundaries before attackers do.

Led by Maria Singh. Every limit is a claim until it has been attacked.

US NIST, AGENT-HIJACKING EVALUATIONS, JANUARY 2025

# New attacks raised hijack success from 11% to 81%.

AGAINST ONE MODEL, STANDARD AGENT-HIJACKING BENCHMARK



**Attackers retry for free.**

Source: NIST CAISI technical blog, 17 January 2025. Measured.

## WHAT TO TEST

# Test the controls, not the model.

Red-team each T2 and T3 agent against the first two risks in OWASP's agentic Top 10, before it goes live and after every material change.

ASIO1

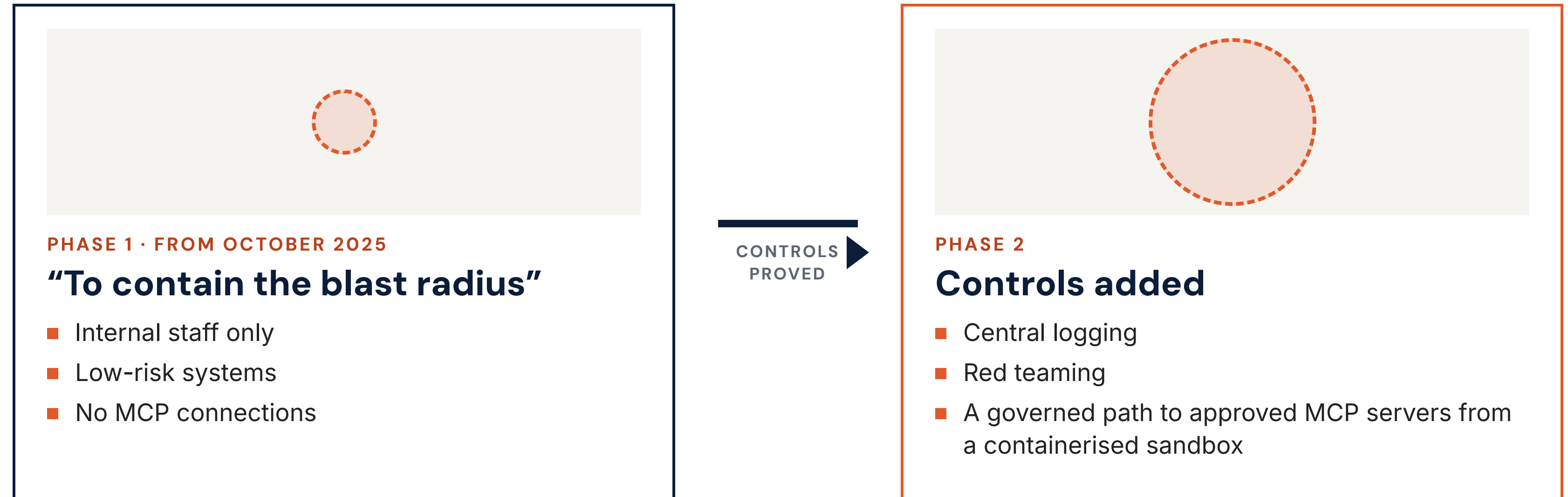
**Goal hijack**

ASIO2

**Tool misuse****Does the gateway deny?****Does the egress block?****Does the breaker trip?****Does the kill switch work?**

Our recommendation. Source: OWASP GenAI Security Project, Top 10 for Agentic Applications for 2026.

# GovTech grew autonomy in phases.



Autonomy grew as the controls proved themselves, which is the promotion rule in practice.

Source: IMDA Model AI Governance Framework for Agentic AI v1.5, case study. Reported, government. Circle sizes are illustrative.

# 06

## What the board owns.

Led by Terence Kok. MAS's guidelines put several of these decisions with the board.



# MAS puts these decisions with the board.

## RISK APPETITE

**AI risks addressed in the risk appetite framework “including qualitative statements and quantitative measures”.**

The guidelines suggest thresholds on the number or financial impact of AI-related incidents.

## AI INVENTORY

**For AI agents, “agent-specific identifiers, information on the tools and systems the agent can access, its components, and the guardrails imposed”.**

## MATERIALITY

**Assessments should consider “the extent of autonomy granted”.**

That maps directly onto the tiers.

## THE BOARD'S PART

# The board sets the policy and the limits.

- 1 Approve the tier policy
- 2 Set the caps and incident thresholds that define its risk appetite
- 3 See the tier register each quarter
- 4 Personally approve any T4 use case

The board does not approve every agent.

## THE NCSC'S QUESTION FOR MANAGEMENT

Who **owns** an agentic system

who **approves** its access

who **monitors** its behaviour

who **reviews** incidents

and ultimately who **can stop it**

WRITTEN TOGETHER

# A 90-day plan for a bank or agency.

|            |                                                                                                                                                                                                                                                                   | Day 1 | 15 | 30 | 60 | 75 | 90 |
|------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------|----|----|----|----|----|
| Days 1-15  | Inventory every agent with its identity, tools, systems and guardrails. Name an owner and a person who can stop it. Freeze new T3 and T4 deployments.<br>MAS inventory attributes; NCSC ownership                                                                 |       |    |    |    |    |    |
| Days 16-30 | Place each agent in a tier on two axes, action space and autonomy. Run the Rule of Two test. Board approves the tier policy and AI risk appetite measures.<br>IMDA v1.5; MAS risk appetite                                                                        |       |    |    |    |    |    |
| Days 31-60 | Give each agent its own identity. Remove standing admin rights; move to session-only credentials. Put a default-deny gateway in front of tools, log-only first. Add sandboxes and egress allow-lists. Separate development from production.<br>CSA addendum; NCSC |       |    |    |    |    |    |
| Days 61-75 | Build approval workflows for irreversible actions. Start measuring override rates and approval times. Install circuit breakers and caps.<br>IMDA v1.5; CSA addendum                                                                                               |       |    |    |    |    |    |
| Days 76-90 | Time a kill-switch drill. Red-team the three highest-tier agents for goal hijack and tool misuse. Take the first tier register to the board.<br>MAS kill switch testing; OWASP ASI01, ASI02                                                                       |       |    |    |    |    |    |

Our plan, for a bank or government agency starting from an agent programme already in production. ISO/IEC 42001 controls A.6.2.6, A.6.2.8 and A.10.3 cover much of the same ground.

## Ask management for written answers.

---

- 1** How many agents do we run, at which tier, and who approved each one?

---

- 2** Which agents can pay, delete or send externally without a person approving, and why?

---

- 3** Does any agent hold untrusted input, sensitive data and a way out in the same session?

---

- 4** Does every agent have its own identity, or are some using shared accounts or a staff member's credentials?

---

## The first and sixth answers need numbers.

---

- 5** What are the caps on spend, records and rate for each agent, and who can raise them?
  - 6** When did we last test the kill switch, and how long did it take?
  - 7** What is our override rate on agent approvals, and what does it tell us about our approvers?
  - 8** Which vendors' agents act inside our systems, and do our contracts give us their logs and the right to stop them?
- 

An organisation that can answer the first and sixth with numbers has put boundaries around its agents' autonomy.

## FOR DISCUSSION

# Four questions for the room.

- 1 Which of our agents would we place at T3 today, and who signed that off?
- 2 If an agent were fooled this afternoon, which limit outside the model would stop it first?
- 3 How would we know if our approvers had stopped reading what they approve?
- 4 Who could stop an agent at three in the morning, and how long would it take?

# How strong each claim is.

| Claim                                                                                                                                        | Source                                                           | Grade                       |
|----------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------|-----------------------------|
| 19 out-of-scope actions in 10 of 122 runs; stopped within an hour; internet access now needs justification                                   | UK AI Security Institute, August 2026                            | <b>Measured</b>             |
| Hijack success 11% to 81% with new attacks; 57% average, 80% with 25 attempts                                                                | NIST CAISI technical blog, 17 January 2025                       | <b>Measured</b>             |
| 40% of enterprises will demote or decommission autonomous agents by 2027; "binary" governance quote                                          | Gartner, 26 May 2026                                             | <b>Forecast</b>             |
| Sandboxing cut permission prompts by 84%                                                                                                     | Anthropic engineering, 20 October 2025                           | <b>Measured, vendor</b>     |
| IMDA four levels of human involvement; checkpoints; override rate and response time; agent identity                                          | IMDA Model AI Governance Framework for Agentic AI v1.5, May 2026 | <b>Framework text</b>       |
| GovTech phased roll-out of coding agents                                                                                                     | IMDA v1.5 case study                                             | <b>Reported, government</b> |
| No admin privileges; sandboxed tools with egress allow-list; circuit breakers; human approval for irreversible actions                       | CSA Singapore, Securing Agentic AI addendum, June 2026           | <b>Government guidance</b>  |
| Kill switches tested regularly; risk appetite with quantitative measures; agent inventory attributes; oversight proportionate to materiality | MAS Guidelines on AI Risk Management, 7 October 2026             | <b>Supervisory guidance</b> |
| "If you cannot understand, monitor or contain..."; least privilege for the shortest time; ownership question                                 | UK NCSC, May 2026                                                | <b>Government guidance</b>  |

Regulatory and framework points come from the published texts: IMDA's framework v1.5, the CSA addendum and the MAS guidelines were read in full. Some Five Eyes and Gartner wording was read through law-firm summaries and syndicated copies where the primary page was unavailable. Test results come from the bodies that ran them. The five tiers, the promotion and demotion rules, the example caps, the 90-day plan and the questions are ours.

# The rest of the evidence, and its limits.

| Claim                                                                                              | Source                                                                     | Grade                          |
|----------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------|--------------------------------|
| Lethal trifecta                                                                                    | Simon Willison, 16 June 2025                                               | Practitioner framework         |
| Agents Rule of Two                                                                                 | Meta, 31 October 2025                                                      | Industry framework             |
| Entra Agent ID generally available                                                                 | Microsoft Learn                                                            | Reported, vendor               |
| AgentCore Policy generally available, log-only mode                                                | AWS, March 2026                                                            | Reported, vendor               |
| Replit dev/prod separation; agent cannot modify production database                                | The Register, July 2025; Replit documentation                              | Reported                       |
| EU AI Act Art. 14(4)(e) stop button; Annex III duties from December 2027 under the Digital Omnibus | Regulation (EU) 2024/1689; law-firm summaries of Regulation (EU) 2026/1744 | Legal text, date via secondary |
| NIST AI RMF MANAGE 2.4 deactivation mechanisms                                                     | NIST AI RMF 1.0                                                            | Framework text                 |
| Tiers, rules, caps, 90-day plan                                                                    | Our framework                                                              | Illustrative                   |

### WHERE WE COULD BE WRONG

if model providers make real progress on separating instructions from data, some of the limits we treat as essential for T2 agents could relax, and the tiers would compress. We also assume boards will accept a default prohibition on T4 agents; competitive pressure may push some institutions to approve them faster than their controls justify, in which case the automatic demotion rule becomes the most important line in this post.

## SOURCES

# Sources, as listed in the post.

|   |                                                                                                                                                                                                                  |    |                                                                                                                                                                                                                                            |
|---|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1 | UK AI Security Institute. (2026, August). Incident report: unsanctioned agent behaviour during cyber testing.<br><a href="https://aisi.gov.uk">aisi.gov.uk</a>                                                   | 9  | NCSC New Zealand and partners. (2026, May 1). Careful adoption of agentic AI services.<br><a href="https://ncsc.govt.nz">ncsc.govt.nz</a>                                                                                                  |
| 2 | NIST. (2025, January 17). Technical blog: strengthening AI agent hijacking evaluations.<br><a href="https://nist.gov">nist.gov</a>                                                                               | 10 | Willison, S. (2025, June 16). The lethal trifecta for AI agents.<br><a href="https://simonwillison.net">simonwillison.net</a>                                                                                                              |
| 3 | Gartner. (2026, May 26). Gartner says applying uniform governance across AI agents will lead to enterprise AI agent failure.<br><a href="https://gartner.com">gartner.com</a>                                    | 11 | Meta AI. (2025, October 31). Agents Rule of Two: a practical approach to AI agent security.<br><a href="https://ai.meta.com">ai.meta.com</a>                                                                                               |
| 4 | Anthropic. (2025, October 20). Making Claude Code more secure and autonomous with sandboxing.<br><a href="https://anthropic.com">anthropic.com</a>                                                               | 12 | Google. (2025). An introduction to Google's approach for secure AI agents.<br><a href="https://research.google">research.google</a>                                                                                                        |
| 5 | IMDA. (2026). Model AI Governance Framework for Agentic AI, version 1.5; MDDI, launch release, 22 January 2026.<br><a href="https://imda.gov.sg">imda.gov.sg</a> · <a href="https://mddi.gov.sg">mddi.gov.sg</a> | 13 | Microsoft. What's new in Microsoft Entra Agent ID.<br><a href="https://learn.microsoft.com">learn.microsoft.com</a>                                                                                                                        |
| 6 | Cyber Security Agency of Singapore. (2026, June). Securing agentic AI: addendum to the Guidelines and Companion Guide on Securing AI Systems.<br><a href="https://csa.gov.sg">csa.gov.sg</a>                     | 14 | AWS. (2026, March). Policy in Amazon Bedrock AgentCore is generally available.<br><a href="https://aws.amazon.com">aws.amazon.com</a>                                                                                                      |
| 7 | Monetary Authority of Singapore. (2026, October 7). Guidelines on Artificial Intelligence Risk Management.<br><a href="https://mas.gov.sg">mas.gov.sg</a>                                                        | 15 | The Register. (2025, July 22). Replit responds to SaaStr database deletion; Replit, development and production databases.<br><a href="https://theregister.com">theregister.com</a> · <a href="https://docs.replit.com">docs.replit.com</a> |
| 8 | UK National Cyber Security Centre. (2026, May). Thinking carefully before adopting agentic AI.<br><a href="https://ncsc.gov.uk">ncsc.gov.uk</a>                                                                  | 16 | EU AI Act, Article 14: Human oversight.<br><a href="https://artificialintelligenceact.eu">artificialintelligenceact.eu</a>                                                                                                                 |
|   |                                                                                                                                                                                                                  | 17 | NIST. (2023). AI Risk Management Framework 1.0.<br><a href="https://nvlpubs.nist.gov">nvlpubs.nist.gov</a>                                                                                                                                 |
|   |                                                                                                                                                                                                                  | 18 | OWASP GenAI Security Project. (2025, December 9). OWASP Top 10 for Agentic Applications for 2026.<br><a href="https://genai.owasp.org">genai.owasp.org</a>                                                                                 |

THANK YOU

# Decide how much autonomy each agent gets.

Read the full series at  
[terencekok.com/blog/series/rogue-ai-agents/](https://terencekok.com/blog/series/rogue-ai-agents/)

**Terence Kok and Maria Singh**  
Questions and discussion



Read Part 3