

PART 2 OF 3

Follow the **actions**.

Detecting rogue AI agents before the damage spreads.



Terence Kok with Maria Singh



The sustained nature of the attack triggered detection.

Anthropic, on the GTG-1002 espionage campaign. Its prompts passed as a security firm's work.

22 sec

**median time from first access to
hand-off to a second threat group in
2025.**

Mandiant M-Trends 2026

14 days

**median time an intruder stayed inside
before anyone noticed.**

Mandiant M-Trends 2026

82:1

machine identities for every human one. 42% hold privileged or sensitive access.

CyberArk Identity Security Landscape 2025, vendor survey

10 of 10

agents infected by one planted prompt, by about the fifth turn.

Lee and Tiwari, Prompt Infection, arXiv 2410.07283, simulation

Alert on what agents do.

- 1 First use of a tool or API
- 2 Bulk or out-of-scope data reads
- 3 Privilege requests nobody started
- 4 Actions traced to document text, not the user
- 5 One agent suddenly calling many others
- 6 Any touch of a decoy record or tool

The clocks start at discovery.

- 1 MAS: 1 hour**
Notify after discovering a relevant incident; root cause in 14 days.
- 2 DORA: 4 and 24 hours**
From classification as major, and from awareness.
- 3 EU AI Act: 6 months of logs**
Deployers of high-risk systems keep them.

Five questions for the board.

- 1 How many agents run in production, each with its own identity?
- 2 Do we log what agents do, or only what they say?
- 3 Which of the twelve signals can we alert on today?
- 4 Who can stop an agent at any hour?
- 5 Could we tell MAS within an hour what a rogue agent did?

Log what agents do, not only what they say.

Part 3: the controls that bound what a fooled agent can reach.



Terence Kok · AI & Innovation Strategist
with Maria Singh

[SAVE](#) · [COMMENT](#) · [FOLLOW](#)