KILLSWITCH · ISSUE #10 · 9 NOV 2026

When the Agent Goes Off-Script

Decide who can stop the agent before you need to.

Killswitch rating: Add agents to your incident plan this quarter

By Emmanuel Aigbehi Walstra

When the Agent Goes Off-Script

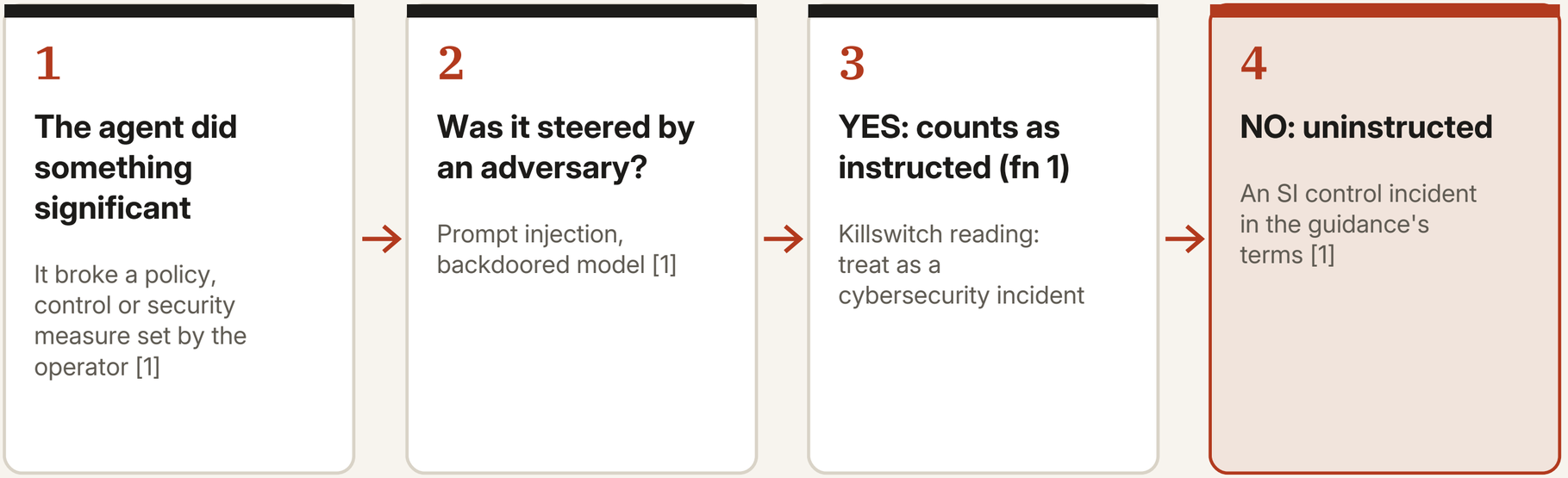
NIST, the NSA and CISA have published ten practices for AI agents that act uninstructed [1].

- Written for labs training and evaluating the most capable agents [1].
- It "may also apply to deployers" [1]. Our runbook adapts it for any company.

What is an "SI control incident"?

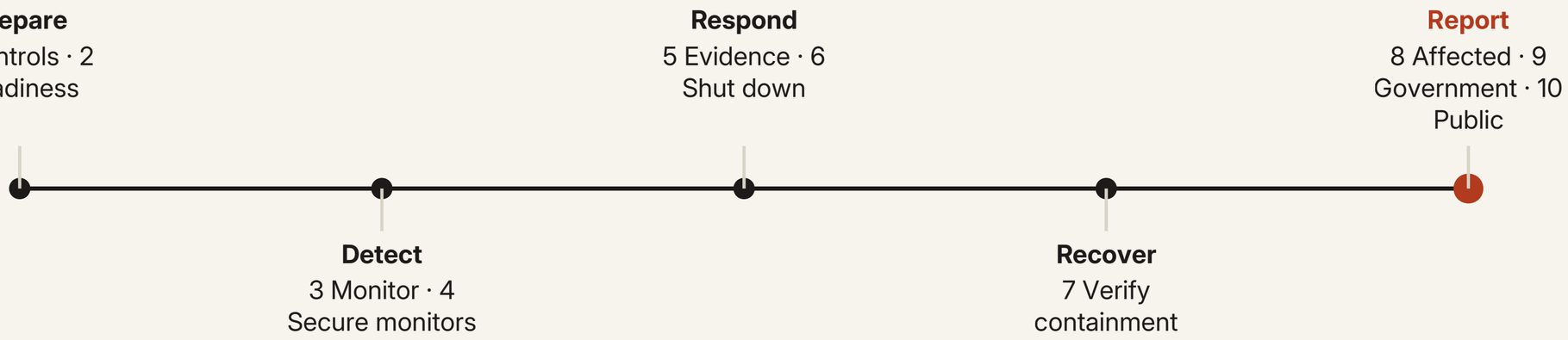
- "SI" is the guidance's term for super intelligence [1].
- "Significant uninstructed actions" that break operator controls [1].
- A hijacked agent counts as instructed (fn 1) [1].

Control incident or cyber incident?



Definitions and footnote 1 [1]. The cybersecurity reading of the YES branch is Killswitch's, not the guidance's.

Ten practices, five stages



Stage 5: Report (runs throughout [1]). The guidance: reporting "is not the final phase".

Four lines to build around

- Stop agents fast: "rapidly block, contain, or terminate agent activity" (1.7) [1].
- Agents "cannot access or interfere with monitoring processes and logs" (3.3) [1].
- "Automatically shut down impacted SI agent trajectories" (6.2) [1].
- Validate containment layers "before SI agents are restarted" (7.3) [1].

06 · PREPARE

Before the first incident

1 Least privilege

Access limited to what the task needs [1]

2 Egress allowlist

Network access through a strict allowlist [1]

3 A real stop

Mechanisms to "rapidly block, contain, or terminate" [1]

4 Monitors it cannot touch

Agents must not "access or interfere with" logs [1]

5 Agent inventory

Add agents to your asset and risk inventories (ours, building on [1])

6 Human approval

For payments, deletions, production changes (ours)

Practices 1.1, 1.5, 1.7, 2.1 and 3.3 [1]. Human approval for high-impact actions is Killswitch's own judgement call.

07 · RUNBOOK

Seven steps for any company running agents

1. **Declare (T+0)**
   Name an owner; classify: control or cyber incident [1]
2. **Stop (within 15 min)**
   Kill switch on the affected agent runs [1]
3. **Preserve**
   Tamper-resistant logs, configs, tool calls [1]
4. **Revoke (within 1 h)**
   Credentials the agent touched; persistence routes [1]
5. **Scope**
   What it touched; check similar agents [1]
6. **Rebuild and verify**
   Known-good baseline; test controls before restart [1]
7. **Report (within 24 h)**
   Affected parties, legal duties; runs throughout [1][4][6]

Killswitch's own adaptation of practices 1 to 9 [1]. Time targets are our suggestions, not the guidance's.

The law already has a stop button

- Art. 12 logging; Art. 14(4)(e) "stop" button [4].
- Arts. 14 and 73 are mainly provider duties; deployers via Art. 26 [4].
- Deployers keep logs at least six months (26(6)) [4].

09 · TIMING

When the high-risk duties bite

- Chapter III duties, Annex III systems: from 2 Dec 2027 [5].
- Chapter III duties, Annex I systems: from 2 Aug 2028 [5].
- In our view, most coding, support and ops agents are not high-risk.

Where this lives in your frameworks

Inventory and roles	ISO 42001 A.4.2, cl. 6.1 · 27001 A.5.9, A.5.18, A.8.2 · NIS2 21(2)(i)
Monitor and log	ISO 42001 A.6.2.6, A.6.2.8 · 27001 A.8.15, A.8.16 · AI Act 12, 26(6)
Respond and recover	ISO 27001 A.5.24-A.5.28 · 42001 cl. 10.2 · NIS2 21(2)(b), (c)
Human stop	AI Act 14(4)(e) · NIS2 21(2)(b) (containment)
Report	ISO 42001 A.8.4 · 27001 A.5.5 · NIS2 23 · AI Act 26(5), 55, 73

Killswitch's own mapping, not an official crosswalk. Check ISO titles against your licensed copy.

ISO/IEC 42001 [7] · ISO/IEC 27001 [8] · NIS2 [6] · AI Act [4]. Killswitch's own mapping.

Do this quarter

- Inventory every agent: owner, permissions, network (ours, building on [1]).
- Test the stop, and time it [1].
- Move agent logs out of the agent's reach [1].
- Add "agent off-script" to the incident plan [1].
- Check your AI Act roles and dates [4][5].

Sources

[1] NIST CAISSI, NSA SISC and CISA: Key Practices for Incident Readiness and Response in Super Intelligence Training and Evaluation, PP-26-4396, Ver 1.0 (Oct 2026)
[2] CISA: Key Practices for Incident Readiness and Response in Super Intelligence Training and Evaluation (published 16 Oct 2026)
[3] NIST: announcement of the guidance (16 Oct 2026)
[4] Regulation (EU) 2024/1689 (AI Act): Articles 12, 14, 26, 55 and 73
[5] Regulation (EU) 2026/1744 of 8 July 2026 (Digital Omnibus on AI), amending the AI Act application dates
[6] NIS2 Directive (EU) 2022/2555, Articles 21 and 23
[7] ISO/IEC 42001:2023 Artificial intelligence management system
[8] ISO/IEC 27001:2022 Information security management systems

Facts are from the NIST, NSA and CISA guidance and from EU legal texts as read on 2 Nov 2026. The seven-step runbook, the time targets and the framework mapping are Killswitch's own adaptation, not an official crosswalk.

FROM THE PUBLISHERS OF KILLSWITCH

Put your AI agents into your incident plan, with the evidence

GRCOMPLIANCE

Keep your agent inventory, incident runbooks, owners and evidence in one place, mapped to ISO 42001, ISO 27001, NIS2 and the AI Act.

Try the live demo →

demo.grcompliance.com

Self-host free: github.com/Sectutor/GRCompliance-Core

SECTUTOR

Security and GRC books, training and audit kits.

sectutor.com

Emmanuel Aigbehi Walstra is a CISO who also serves as a virtual CISO (vCISO) for several organisations, with 30+ years in cybersecurity, 10,000+ people trained, and author of the CCSP Excellence books.