

How AI agents turn low-rated security findings into a **serious risk**

Why findings scored one at a time miss the attack path, and who should own it.

Terence Kok with Maria Singh

How a findings report is built

- 1 Each finding is scored on its own
 - 2 Each is assigned to the team that owns that system
 - 3 Lows are deferred or accepted, one by one
-

No step asks whether one team's Low makes another team's Low reachable. Attackers ask exactly that.

An accounts payable agent with three Low findings

A Untrusted content

Supplier emails reach the agent unchecked

Low · AI product team

B Private data

Agent account reads the whole finance library

Low · Identity and access team

C External send

Agent emails attachments out without approval

Low · Finance operations

Chained severity: High. Path owner: none.

3

conditions that make an AI agent exploitable together

Access to private data. Exposure to untrusted content. A way to send data out. Each one tends to score Low when reviewed alone.

Simon Willison, "the lethal trifecta", June 2025

4

filter bypasses chained in EchoLeak, a zero-click Microsoft 365 Copilot flaw

Each safeguard worked most of the time. Together they took data from one crafted email to the attacker. Microsoft scored the chain 9.3, Critical.

CVE-2025-32711; Reddy and Gujral, 2025. Fixed May 2025, no evidence of exploitation.



code flaws in the GitHub MCP server behind a private-repository leak

A public issue hijacked a coding agent whose credential reached every repository. Many users had set approvals to "Always Allow".

Invariant Labs, 26 May 2025

Score the chain as well as the parts

- 1 Identify which findings combine into one attack
 - 2 Take the easiest way in and the worst impact
 - 3 Report the chained score next to the individual ones
-

FIRST, CVSS v4.0 User Guide, section on vulnerability chaining

Three controls that break the chain

B Least privilege

Scope the agent to the records the task needs

OWASP LLM06:2025; Agentic Top 10, ASIO3

C Approval for consequential actions

Tie it to the action class, keep it rare enough to read

OWASP LLM06:2025

A Adversarial testing

Test the assembled system, inbox to outbound

OWASP GenAI Red Teaming Guide

Five changes so someone owns the path

- 1 Name a path owner for every AI agent
 - 2 Keep a one-page capability map per agent
 - 3 Run a chain review before closing any finding
 - 4 Move acceptance of chained Lows to the path owner
 - 5 Retest the whole path after every material change
-



Every finding had an owner. The attack had none.

Ask how many of your AI agents hold all three conditions, and who answers for each one.



Terence Kok · AI & Innovation Strategist

SAVE · COMMENT · FOLLOW