The sentence that hijacks an agent is a polite line in a document you trusted.

Dr Peter McCann Strain

CTO, DPhil/PhD in AI from Oxford University

Swipe ››

THE INCIDENT PATTERN

Indirect prompt injection (instructions hidden inside content the agent reads) arrives disguised as a pull request, web page, ticket, email or tool response. Models do not natively enforce the boundary between 'the user instructed me' and 'a stranger wrote this in content I was asked to inspect'.

— THE REFRAME

The stream is the boundary.

THE OLD QUESTION

Will the model obey the developer?

↓ ↓ ↓

THE QUESTION THAT HOLDS UP

Can the runtime prove which speaker may steer which action?

—— WHAT THE RECORD SHOWS

1 PR review

Invariant Labs published a working GitHub MCP exploit in 2025 where a single embedded sentence in a public issue could exfiltrate a private repository's contents from a single PR review.

SOURCE

Invariant Labs, GitHub MCP vulnerability disclosure (invariantlabs.ai/blog/mcp-github-vulnerability); with Greshake et al. (2023), the UK NCSC's 'Prompt injection is not SQL injection', OWASP's agentic taxonomy, and Simon Willison's lethal trifecta.

FAILURE CHAIN

Run four runtime corners before an agent reads outside content.

01 Apply **source labelling**: mark every span by trust before the model reads it.

02 Enforce **trifecta refusal**: the trifecta is private data, untrusted text and an outbound channel; never combine all three.

03 Hold **capability gating**: trusted code chooses the next tool, not untrusted text.

04 Run **output checking**: inspect what leaves the boundary before it lands.

— THE ARTIFACT

Four corners of runtime defence.

Source labels

What trust zone is this text from?

Trifecta break

Remove one path to harm

Capability flow

Gate what authority can move

Output gating

Check before anything leaves

Four corners to run before an agent reads outside content. Miss one, and a sentence can steer the loop.

— DO THIS AFTER THE NEXT INCIDENT

Take one agent workflow that reads outside content this week. List its input streams. Check whether the lowest-trust stream shares a session with private data and an outbound channel. If yes, pick the first of the four corners to build.

FAILURE MODE TO AVOID

Adding more guardrails to the prompt. Stronger incantation doesn't create a data-command boundary. The defence lives at the tool gateway, context builder, egress policy and approval screen, not in the system prompt.

— USE THE FULL POSTMORTEM

The Sentence That Owns the Agent

Read the full essay — the argument, the sources, the figures and a reader-ready working artifact.
Substack petermccannstrain.substack.com · Medium @peter.mccann.strain · LinkedIn peter-strain-dphil-15a607128
New essays twice weekly, 2 June – 21 July 2026.

Next: E12 — The Cheapest Token

— THE STACK SO FAR

E11 · Essay 11 of 22 complete · Arc III: Runtime control

YOU JUST ADDED

The Four-Corner Runtime Defence

STACK LAYER LIT UP

Tools / Permissions / Checks

YOU CAN NOW ASK

defend against runtime authority capture.

NEXT

E12 asks why cost is the budgetary form of delegated authority.

Dr Peter McCann Strain

CTO, DPhil/PhD in AI from Oxford University

I build production AI systems and write about making agentic AI useful, inspectable, governable and safe enough for real work.

Follow on Substack for the full 22-essay series
petermccannstrain.substack.com