



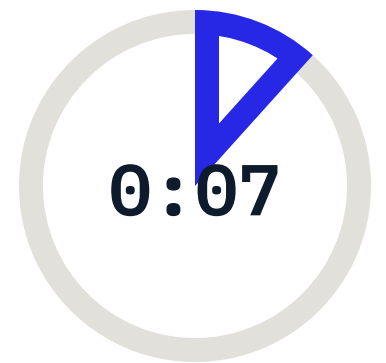
THE PATTERN BEHIND AGENT ATTACKS

The lethal trifecta.

Three ordinary capabilities. Combined, they can turn
an AI agent into an open door.

Manageable alone. Dangerous together.

ONE IDEA • 60 SECONDS • EVERY WEEK



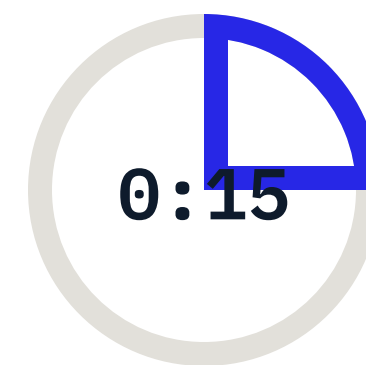
THE PATTERN

Three capabilities. One dangerous combination.

Coined by security researcher Simon Willison in June 2025, the lethal trifecta describes an agent that combines:

- 1 PRIVATE DATA ACCESS**
it can read your files, email, or database
- 2 UNTRUSTED CONTENT**
it processes content controlled by someone else
- 3 EXTERNAL COMMUNICATION**
it can send information outside the system

Any one can be useful. The risk emerges when all three meet.

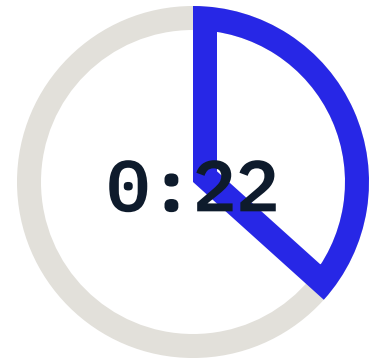


WHY IT'S DANGEROUS

The model can't reliably distinguish data from instructions.

An agent may read an email, web page, document, or support ticket containing malicious instructions. Those instructions enter the model's context alongside trusted instructions. The model may then act on them.

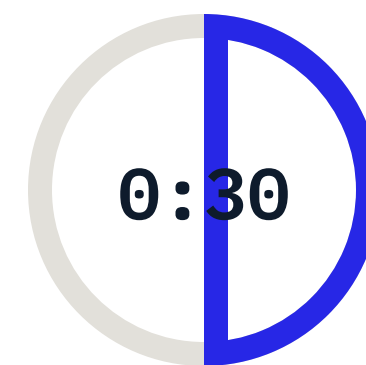
Prompt injection turns untrusted content into potential instructions.



HOW THE ATTACK RUNS

- 1 · Attacker puts malicious content in front of the agent
- 2 · The content tells the agent to retrieve private information
- 3 · The agent follows the injected instruction
- 4 · An external channel carries the data out

No traditional exploit is necessary. The risk comes from the combination of capabilities.



NOT THEORETICAL

We've seen this pattern in production systems.

Researchers have repeatedly demonstrated prompt-injection attacks where untrusted content causes AI systems to access or expose sensitive information.

EMAIL

malicious instructions
embedded in messages

DOCUMENTS

hidden instructions in files

WEB CONTENT

attacker-controlled pages

TOOLS / MCP

combinations that expose
data and provide an
exfiltration path

Different products. Same underlying structure.



BREAK THE CHAIN

You don't always need all three.

The strongest mitigation is to remove a leg of the trifecta when you can:

NO PRIVATE DATA

Limit what the agent can access.

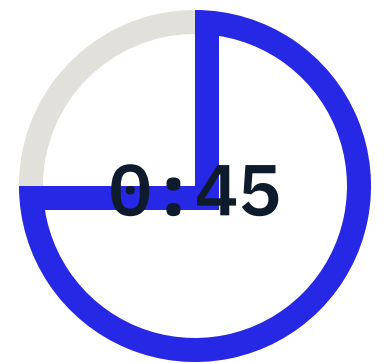
NO UNTRUSTED CONTENT

Don't expose the agent to attacker-controlled input.

NO EXTERNAL COMMUNICATION

Block outbound channels.

If you can remove one leg, the trifecta is broken.



IF YOU CAN'T REMOVE A LEG Then constrain all three.

SCOPE DATA

Per-task access, not standing access to everything.

LIMIT TOOLS

Expose only the capabilities the task requires.

CONTROL OUTPUT CHANNELS

Restrict where data can leave.

REQUIRE APPROVAL

Put a human or policy gate before consequential actions.

LOG & TRACE

Make every sensitive tool call reviewable.

Reduce what the agent can see, do, and send.



TIME 'S UP

Break one leg. Or constrain all three.

Don't rely on the model to recognize the attack. Design the system so the attack can't become an action.

Save this before your next agent gets tool access.

SOURCE

Simon Willison, “The lethal trifecta for AI agents” (June 2025) · simonwillison.net